Geoderma Regional 21 (2020) e00284

Contents lists available at ScienceDirect

Geoderma Regional

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

Evaluating an adaptive sampling algorithm to assist soil survey in New South Wales, Australia

Jingyi Huang [a,b,*], Alex B. McBratney [b], Budiman Minasny [b], Brendan Malone [c,b]

[a] Department of Soil Science, University of Wisconsin-Madison, 1525 Observatory Drive, Madison, WI 53706, USA
[b] Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW 2015, Australia
[c] CSIRO, Agriculture and Food, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 20 February 2020
Received in revised form 17 April 2020
Accepted 23 April 2020

Keywords:
Vertisols
Alfisols
Inceptisols
Entisols
Oxisols

ABSTRACT

Knowledge of the spatial variation of soil is important in modern agricultural management. To attain this knowledge, ground-based samples are required in combination with many ground-based, air-borne and space-borne sensors from the Internet of Things. Compared to traditional grid and simple random sampling that are designed for fixed sensors, adaptive sampling is not well studied. In this study, we propose a prior-based adaptive sampling scheme to collect soil samples for estimation of ground-based Gamma-ray potassium across an 80-ha field in a semi-arid landscape, in New South Wales, Australia. We compare the performance of the sampling algorithm via a linear mixed model between various adaptive sampling schemes with prior information of varying quality (e.g. ground apparent electrical conductivity, air-borne Gamma-ray potassium, and a legacy map of clay content). We also compare the model performance of the adaptive sampling scheme with more conventional grid and simple random sampling schemes. Results show that the adaptive sampling scheme was superior to the grid and simple random sampling schemes in terms of the accuracy of the linear mixed model when the sampling size was small (<15 additional samples) due to the use of prior information. The accuracy of the linear mixed models associated with the adaptive sampling schemes deteriorated when the quality (correlation with the target soil variable) of the prior information decreases. We conclude that the algorithm has the potential to be applied generally for automated adaptive sampling design (e.g., on an autonomous vehicle) when sampling cost is large and travelling time of the sensor is relatively small.

1. Introduction

Knowledge of the spatial variation of soil is important in modern agricultural management (McBratney et al., 2005). To quantify the spatial variation of soil and evaluate the performance of existing soil maps, ground-based soil samples are crucial. However, collection of soil samples in the field can be time-consuming, labour-intensive and expensive. To overcome the limitation of soil sample collection in the field, various ground (Cosh et al., 2004), air-borne (Kramer, 2002; Berni et al., 2009), and space-borne (Hart and Martinez, 2006; Pettorelli et al., 2014) sensors have been used. This is now recognised as part of the Internet of Things (IOT) (Gubbi et al., 2013; Wang et al., 2013; Fang et al., 2014).

Sensors used in soil and environmental studies can be classified into two categories: fixed and adaptive. Fixed sensors refer to sensors that are located on the ground at different stations and measure physical, chemical and biological properties of variables at different time intervals or on the satellites whose orbits are fixed and survey the globe at a specified spatial resolution and revisit time intervals. Once these fixed sensors are installed, the measuring frequencies and spatial resolutions cannot be modified.

Unlike the fixed sensors, adaptive sensors are placed on ground-based mobile sensor systems, vehicles, or aircrafts. Compared to passive sensors, these sensors can be designed to collect observations, proactively, and adjust their survey routes in real-time based on collected data. Successful applications of adaptive sensors include automated driving systems (Kato et al., 2002), cleaning robots (Bartsch et al., 2002; Jones et al., 2005), active learning for text classification (Tong and Koller, 2001) and image retrieval (Tong and Chang, 2001).

Brus and Heuvelink (2007) provided a summary of sampling methods that can be directly used to select the locations of the fixed sensors. In general, these sampling algorithms can be classified as design-based or model-based methods (Brus and De Gruijter, 1997). In the design-based approach, stochasticity is introduced at the stage of sampling and sample locations are selected by a pre-determined random selection procedure. By contrast, model-based methods build models to estimate the probabilities of the sampled target variables as a stochastic process (e.g. presence of variations of soil properties due

[*] Corresponding author at: Department of Soil Science, University of Wisconsin-Madison, 1525 Observatory Drive, Madison, WI 53706, USA.
E-mail address: jhuang426@wisc.edu (J. Huang).

https://doi.org/10.1016/j.geodrs.2020.e00284
2352-0094/Crown Copyright © 2020 Published by Elsevier B.V. All rights reserved.

to soil-forming processes), which is a mathematical abstraction used to describe reality. Biswas and Zhang (2018) provided a summary of the popular methods used in soil mapping, including simple random sampling, grid sampling, cluster sampling, transect sampling, nested sampling, spatial coverage sampling, stratified random sampling, Latin hypercube sampling (LHS) and fuzzy k-means sampling. Most of these classical sampling algorithms select sampling locations before the soil survey starts, and is not affected by the locations and values of the samples. Here, we consider these sampling algorithms as non-adaptive sampling.

Compared to non-adaptive sampling that estimates the probabilistic distribution of sampled variables in space and time (design-based) or reducing prediction errors (model-based) (Webster and Lark, 2012), adaptive sensors can be used with adaptive sampling algorithms to estimate soil properties in a proactive way. Adaptive algorithms, sometimes known as active learning (Cohn et al., 1996; Salganicoff et al., 1996), often comprise objective functions that simultaneously optimise prediction errors and other variables such as survey time (Martinez-Cantin et al., 2007; Kroemer et al., 2010; Kulick et al., 2013). Unlike the non-adaptive sampling, the locations of the samples are determined during the sampling process, where the selection of subsequent samples is based on information obtained from the previous samples so that both the prediction error and survey time can be minimized as the soil survey continues.

With the development of sensing technology and robotics, adaptive sampling has been increasingly used in many fields, such as in agriculture (McBratney et al., 2005; Tokekar et al., 2016), geology (Potts et al., 2015), hydrology (Singh et al., 2014), and environmental sciences (Rahimi et al., 2004). Although different adaptive algorithms have been previously proposed for mapping environmental variables in space with the use of newly collected data as input information for selecting subsequent sampling locations (Flajolet, 1990; Cox, 1999; Marchant and Lark, 2006; Musafer and Thompson, 2016), few algorithms have been designed for minimizing prediction error and sampling costs (e.g. travel and sampling time) at the same time. In addition, previous adaptive sampling algorithms are not able to incorporate prior information such as previous surveys or auxiliary covariates that have different levels of data-quality or correlation to target variables.

The objectives of this study are: 1) to establish an adaptive sampling algorithm that uses prior information for estimation of a target soil variable during the soil survey, and 2) to evaluate the model performance between adaptive sampling with prior information of different quality, grid and simple random sampling algorithms. The hypotheses here are: 1) the adaptive sampling with prior information is superior in efficiency or cost to traditional grid and simple random sampling, and 2) the performance of the adaptive sampling varies with decreasing quality of prior information.

2. Materials and methods

2.1. Study site

The study site is situated at Nowley Farm, on the Liverool Plains region in north-northwest New South Wales, Australia. The study area consists of two agricultural fields totalling 84 ha (Fig. 1a) divided by a road traversing from southeast to northwest. The soils are mostly Black Vertosols (Australian Soil Classification) or Udic Haplusterts (United States Department of Agriculture Soil Taxonomy) (Stockmann et al., 2016). The field has an annual maximum temperature of 24.6 °C and annual minimum of 12.2 °C with 637 mm precipitation. The main land use is pasture.

2.2. Collection of the target soil variable

The target soil variable used in this study was gamma-potassium (Gamma-K), namely, the natural emissions of soil gamma-rays from ^{40}K. Studies have shown that gamma-ray data (e.g. Gamma-K) are related to various soil properties such as soil mineralogy (Wilford et al., 1997), clay content (Wong and Harper, 1999; Pracilio et al., 2006), and soil types (Schuler et al., 2011). Because most gamma-rays were emitted within the top 0.3 m of the soil (Minty, 1997), gamma-ray measurements have been mainly used to infer the spatial variation of the topsoil properties.

Gamma-K data was collected on 27 February 2018 with a Sodium–Iodine crystal (Radiation Solutions Inc., Mississauga, Ontario, Canada). The sensor was mounted on a four-wheel drive vehicle with a Real-Time Kinetic Global Positioning System for geo-reference. Gamma-K measurements were interpolated from 20-m spacing transects onto a 25 m × 25 m grid (Fig. 1b) using ordinary kriging with local exponential variograms and a neighbourhood of 90–100 points. In this study, a grid was made of regularly spaced points and the mapping/interpolation process was performed on these points. This discretisation was used in this study to simplify the demonstration of the different sampling approach. There were 1350 points in the grid across the field. The kriging was carried out in Vesper Software (Minasny et al., 2006). To differentiate this dataset from the airborne Gamma-K data used in the next section, we refer to it as Gamma-K-ground.

It should be noted that we did not select any soil physical, chemical or mineralogical properties as the target soil variables although soil properties varied across the study field. This was because a larger number (>1000) of soil samples and subsequent laboratory analyses of the soil properties was not available. As such, the sensor-based measurement (i.e. Gamma-K-ground) was used to test the proposed sampling algorithm at the field scale.

2.3. Collection of prior information

In this study, three types of ancillary data were selected as prior information for an adaptive sampling algorithm and mapping of the target soil variable (i.e. Gamma-K-ground). The first one was collected using a DUALEM-21S (DUALEM Inc., Ontario, Canada) using the same on-the-go soil sensing system as the gamma-ray survey. The DUALEM-21S measures the ability of the bulk soil to conduct electrical currents, namely, apparent electrical conductivity (EC_a, mS m^{-1}), at different coil arrays via non-invasive electromagnetic induction. The EC_a values from the 1-m perpendicular receiver coil (PRP-1 m) has an effective measurement depth of 0.5 m based on the spacing of the transmitter and receiver coils (DUALEM User Manual, 2010). Because EC_a is a function of soil clay content, moisture, and salinity (Corwin et al., 2003; Corwin and Scudiero, 2019), it can be used as prior information to potentially infer the spatial variation of the top 0.5 m of soil and predict the spatial distribution of Gamma-K-ground data. The PRP-1 m EC_a were similarly interpolated onto the same 25 m × 25 m grid using ordinary kriging with a local exponential variograms with a neighbour of 90–100 points.

The second ancillary data set was Gamma-ray K collected from the airborne surveys by the Australian Department of Mineral Resources, termed Gamma-K-air. Similar to ground-based Gamma-rays, the airborne data also represent the variations in soil properties within the top 0.3 m. The Gamma-K-air dataset used in the study was extracted onto the previous 25 m × 25 m grid with the nearest-neighbour algorithm from a harmonised 100 m × 100 m grid across the Australian continent (Minty et al., 2009).

The third ancillary data set was the average soil clay content at the depth of 0–0.3 m (from Soil and Landscape Grid of Australia). It was calculated using clay content at three depth intervals (i.e. 0–0.05, 0.05–0.15 and 0.15–0.3 m) from a 90 m × 90 m grid (Grundy et al., 2015) and was extracted onto the same 25 m × 25 m grid using the nearest-neighbour algorithm. The clay content data was estimated using the Australian soil site collation database (Searle, 2014) and various environmental covariates with a machine learning algorithm or disaggregating existing soil maps (Grundy et al., 2015). We chose this

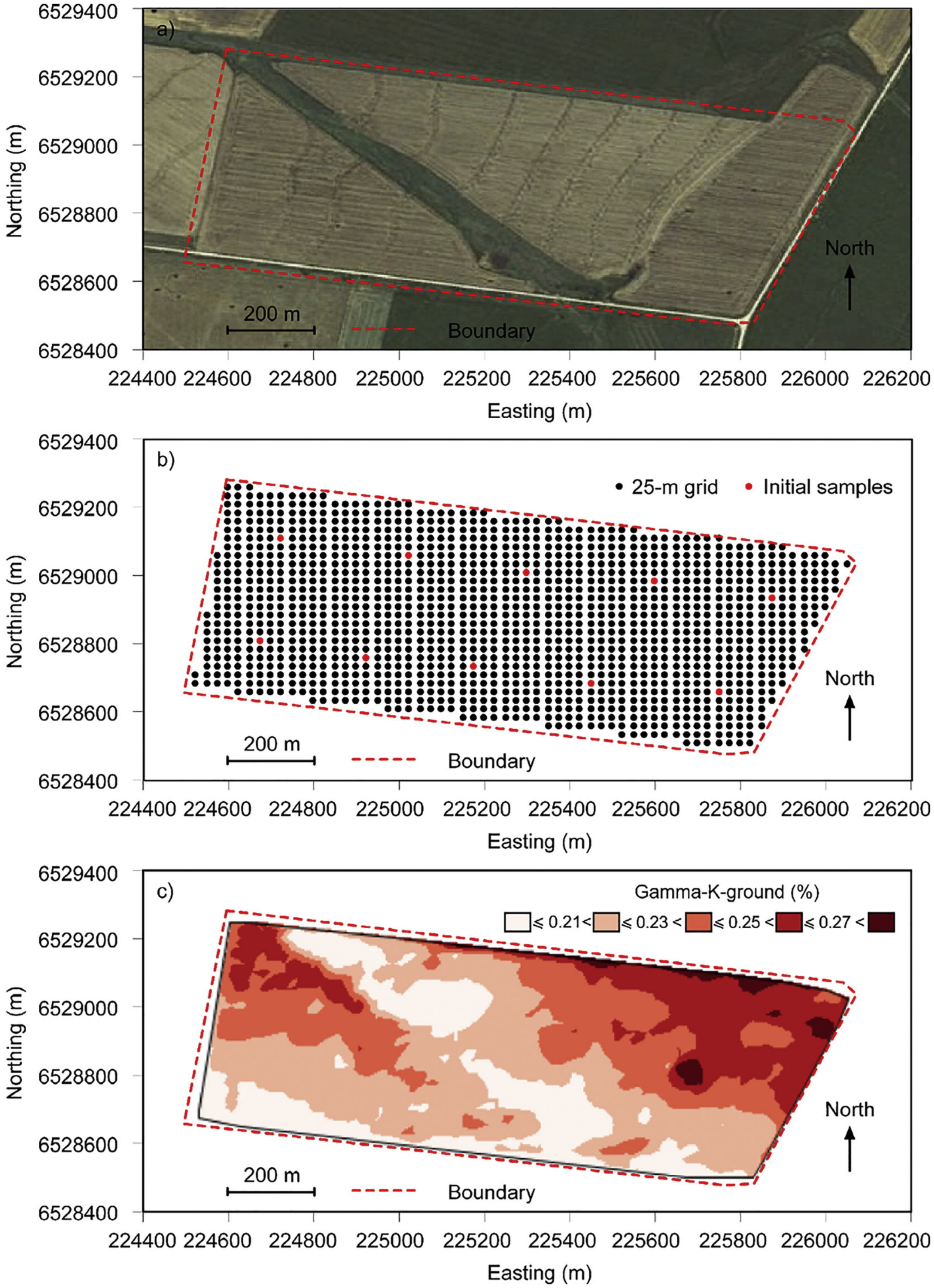

Fig. 1. a) Google Earth imagery, b) locations of the sampling grid and initial samples, and c) contour plot of Gamma-ray potassium measured using the on-the-go soil sensing platform (Gamma-K-ground, %) across the study area the University of Sydney farm, Nowley, New South Wales, Australia. Note: the coordinate system is in Universal Transverse Mercator (UTM), Zone 56S.

depth because it was consistent with the effective measuring depth of Gamma-ray data (Minty, 1997).

2.4. An adaptive sampling algorithm

We used different sampling approaches to sample the Gamma-K data from the grid of Gamma-K-ground (assuming no measurement error) and estimate the target soil variable (i.e. Gamma-K-ground) across the grid based on spatial models established using the samples independently or in combination with prior information. The flowchart of the algorithm is provided in Fig. 2.

First, we started with 10 initial soil samples observed across the study area. These 10 samples were chosen using the k-means clustering of the X- and Y- coordinates of data points from the 25 m × 25 m grid

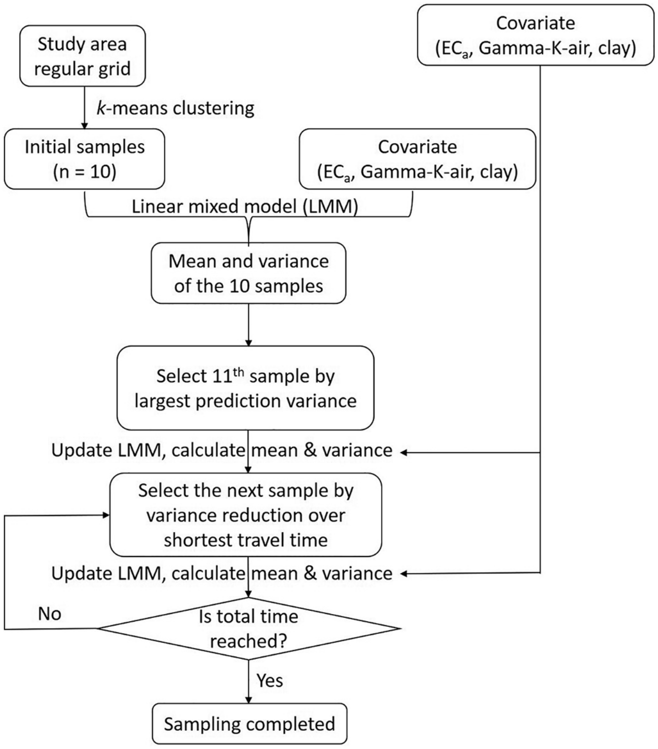

Fig. 2. Flowchart of the adaptive prior-based sampling scheme.

and the centres of the 10 clusters were used as the first 10 initial sampling locations. The k-means clustering of the X- and Y- coordinates enabled the determination of the initial sampling locations using a regular grid sampling approach to cover the field as evenly as possible. The Gamma-K-ground measurements at these 10 sampling locations were stored for the next step.

Second, the algorithm attempted to calculate the spatial distribution of the mean and variance of the target soil variable (i.e. Gamma-K-ground) using 10 initial soil samples and prior information from each of the ancillary data (i.e. PRP-1 m EC_a, Gamma-K-air, or clay content) as a covariate in turn. A linear mixed model (LMM) was used to predict the spatial distribution of the soil variable and can be described using the following equation:

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ is the target soil variable, $\mathbf{X}$ are the covariates (i.e. PRP-1 m EC_a, Gamma-K-air, or clay content) at sampling locations. $\boldsymbol{\beta}$ represents the coefficients to be fitted and $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ represent spatially correlated errors (Lark et al., 2006). The parameter fitting was carried out using the residual maximum likelihood with the geoR package (Ribeiro and Diggle, 2001). Once the parameters of the LMMs were determined, the mean and variance of prediction of Gamma-K-ground were calculated across the 25 m × 25 m grid using gstat package in the R software (Pebesma 2004; Lark and Cullis 2004).

Third, the adaptive sampling algorithm aims at determining the next sampling location (i.e. the 11th sample). This was based on the prediction variance (Var) of the universal (LMM) kriging, which was the sum of the regression variance and the kriging variance. If the covariate is strongly/moderately correlated with the model response, the prediction is mainly based on the regression and the kriging variance will be small. Otherwise, the prediction will be mainly based on kriging, and the kriging variance will be large. The algorithm selects the location with the largest Var as the next sampling location (i.e. the 11th sample). Similarly, the Gamma-K-ground measurements at the sampling locations was used as the measured model response. Along with the previous 10 samples, a new LMM was fitted using all the 11 samples and the mean and variance of prediction were updated across the grid using universal kriging.

Next, the algorithm aimed at selecting the subsequent sampling location (i.e. 12th sample). This time, an additional criterion was added, which was the total time taken for the soil sampling (T(sampling)) and travelling from the current sampling location to the next sampling location (T(travelling)). The new objective function (i.e. time-weighted prediction variance, O(Var)) would penalise locations with largest travel time when comparing the updated variance across the field and became:

$$\mathrm{O(Var)} = \frac{\mathrm{Var}}{\mathrm{T(sampling)} + \mathrm{T(travelling)} \times \mathrm{weight}} \tag{2}$$

Similarly, Var was the updated prediction variance from the previous kriging variance; T(sampling) was 5 min, the time needed to take a measurement from the on-the-go soil sensing system; T(travelling) was calculated using a travelling speed of 5 m/s; the weight was 1.2, which was empirically determined to scale and balance T(sampling) and T(travelling). In this study, we assumed that the time required to collect a soil sample and analyse the target variable was in the same order of that needed for the "robot" surveyor to travel from one location to another location.

Based on Eq. (2), the algorithm selected the next sampling location with the largest ratio of the newly updated variance to the sum of sampling and travelling time. This indicated that instead of travelling to the location with the largest variance as before, the "smart autonomous" surveyor from now on would travel to a location with relatively large variance but within a shorter distance to minimise the mapping variance within a given amount of sampling and travelling time. Similarly, the Gamma-K-ground measurements at the sampling locations was used as the measured model response. Then a new LMM was fitted using all the samples and the mean and variance of prediction were updated again. This process continued until the total time limit ran out. In this process, we selected 100 additional samples using the adaptive sampling algorithm and compared the different results by using each of the covariates (i.e. PRP-1 m EC_a, Gamma-K-air, or clay content) individually in turn.

2.5. Regular grid and simple random sampling designs

To compare the performance of the adaptive soil sampling, regular grid sampling and simple random sampling schemes were also used. The grid sampling was carried out with the same approach used to choose the 10 initial sampling locations by clustering the X- and Y- coordinates. First, the total number of additional samples would be calculated that could be collected within the time limit. For example, if a total of 20 additional samples could be collected, the algorithm would re-cluster the X- and Y- coordinates of the whole field into 20 classes. The centroids of the 20 classes would be used as the locations of the new 20 samples. Given that two sets of samples were used, the overall sampling scheme was a mixture of two grid samples. Along with the previously collected 10 samples (see Section 2.4), all the samples collected were used to fit an LMM. Unlike the adaptive sampling that used ancillary data as covariates, only the X- and Y- coordinates were included in the LMM. This was done to remove the trend of the model response. To compare with the adaptive sampling algorithm, we calculated the mean and variance of prediction of Gamma-K-ground across the grid using a series of additional samples (from 1 to 100).

In terms of the simple random sampling, the algorithm first determined how many samples could be collected within the time limit. After that, a simple random sampling scheme generated by the "sample" function of the R package was used to select the

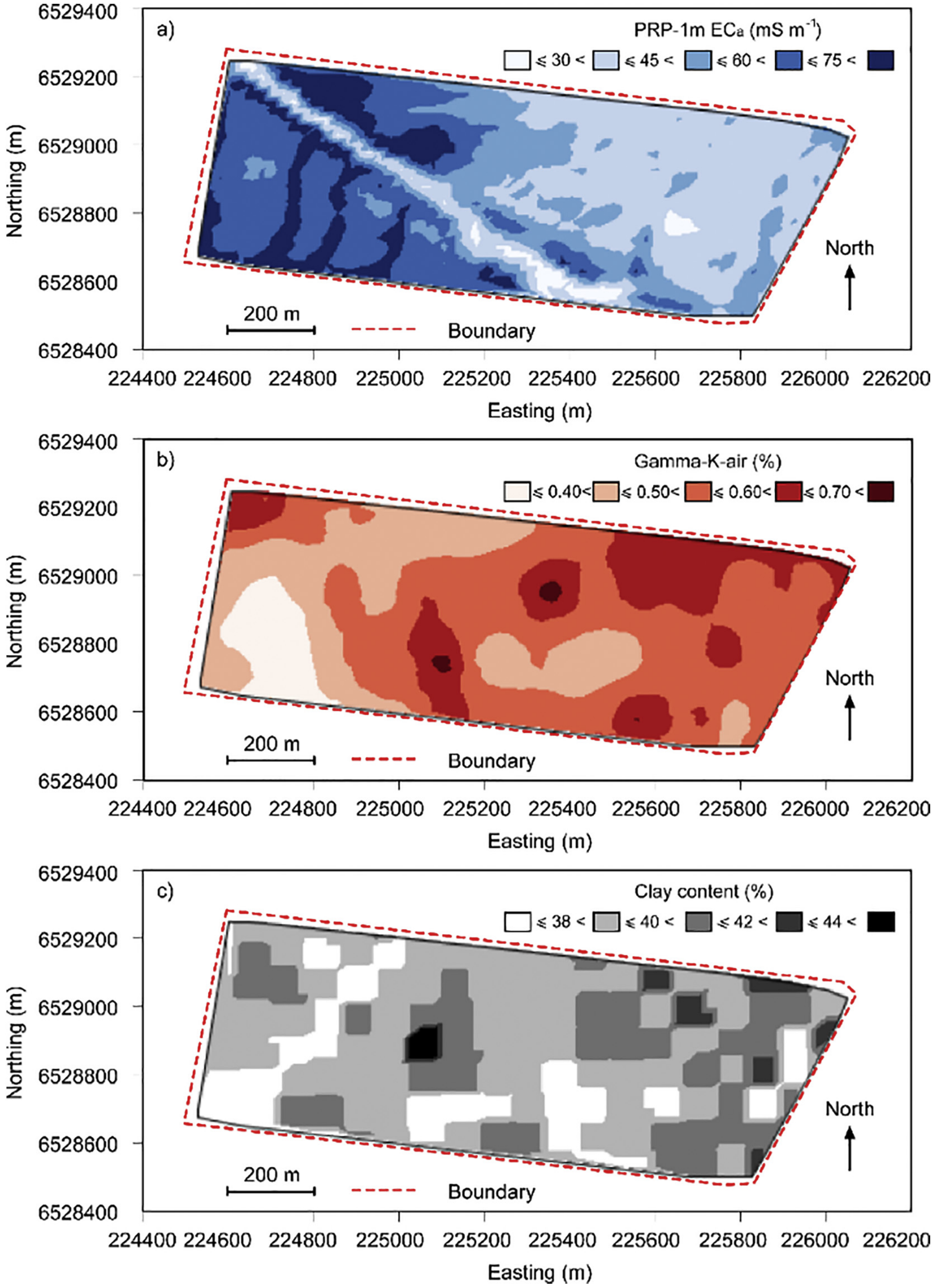

Fig. 3. Contour plots of various ancillary data used as prior information and including; a) apparent electrical conductivity measured by DUALEM 1-m perpendicular coil array (PRP-1 m EC_a, mS m^{-1}), b) gamma-ray potassium obtained from the national airborne radiometric map (Gamma-K-air) (%), and c) average clay content at 0–0.3 m (%) calculated using the Soil and Landscape Grid of Australia.

given number of samples. Along with previously selected 10 samples on the grid, all the samples were used to fit an LMM. Similarly to grid sampling, X- and Y- coordinates were included in the LMM as covariates. To compare with the adaptive sampling algorithm, mean and variance of prediction of Gamma-K-ground across the grid were also calculated using a series of additional samples (from 1 to 100).

2.6. Evaluating the impacts of sampling size

The impacts of sampling sizes on various algorithms were evaluated using several metrics between the Gamma-K-ground interpolated across the grid and predicted Gamma-K-ground by the LMMs. These include R^2, concordance correlation coefficient (Lin 1989), mean error (ME) and root mean square error (RMSE).

Table 1
Statistics of covariates collected across the field and the correlation coefficients.

	N	Min	Mean	Median	Max	SD	Skewness	CV%
Gamma-K-ground (%)	1350	0.16	0.23	0.23	0.29	0.02	0.18	9.5
PRP-1 m EC_a (mS m^{-1})	1350	16.2	54.0	50.7	104.7	16.5	0.38	30.5
Gamma-K-air (%)	1350	0.28	0.53	0.54	0.74	0.08	−0.40	15.4
Clay content (%)	1350	35.7	39.3	39.3	45.8	1.6	0.56	4.0
Pearson's r	Gamma-K-ground	PRP-1 m EC_a	Gamma-K-air	Clay content				
Gamma-K-ground	–							
PRP-1 m EC_a	−0.50	–						
Gamma-K-air	0.39	−0.43	–					
Clay content	0.27	−0.23	0.30	–				

3. Results and discussion

3.1. Spatial distribution of the target soil variable

Fig. 1c shows the pattern of Gamma-K-ground. Note that small Gamma-K-ground values (<0.21%) were identified in the centre of the field, as well as along the southern margin of the field. The variations of Gamma-K values indicate variations of soil properties at the field scale, which may be caused by spatial variations of parent materials during the alluvium deposition process and their different weathering status (Wilford, 1995; Triantafilis et al., 2013).

3.2. Spatial distribution of different covariates

Fig. 3 shows the spatial distributions of EC_a, Gamma-K-air, and clay content. The patterns of these environmental covariates were similar to that of the Gamma-K-ground. As shown in Table 1, EC_a was most strongly correlated with Gamma-K-ground ($r = -0.50$), followed by Gamma-K-air (0.39), and clay content (0.27). The differences in correlation between different covariates with Gamma-K-ground are due to the different measuring depths of the instruments/soil maps as well as the different responses of electrical magnetic fields and gamma-ray emissions to soil properties. Similar correlations were identified in other studies between Gamma-K-ground data with soil clay content (e.g. Petersen et al., 2012; Spadoni and Voltaggio, 2013) and EC_a (e.g. Piikki et al., 2013) although the correlation coefficients varied from site to site. Because of the different correlation coefficients between covariates and Gamma-K-ground, it is expected that LMMs built using different covariates will have different performance in predicting Gamma-K-ground.

3.3. Model performance: adaptive sampling vs. grid sampling and simple random sampling

Fig. 4 shows the model performance for adaptive sampling and the grid sampling. As indicated by coefficient of determination (R^2) and Lin's concordance (Fig. 4a), adaptive sampling schemes with a moderately correlated covariate (EC_a) outperformed grid sampling scheme when the size of additional samples was small (<15). However, when the additional sample size was greater than 15, the grid sampling scheme was better than the adaptive sampling scheme. When the sample size was large (>80), both sampling schemes performed well ($R^2 =$ 0.8, concordance = 0.85). Similar patterns were observed for RMSE between the adaptive sampling scheme and grid sampling scheme. This suggested that in terms of model accuracy, adaptive sampling with a moderately correlated covariate (i.e. EC_a) was superior to grid sampling scheme when the additional sampling size was small (< 15), worse than the grid sampling when the sampling size was intermediate (15–80), and equivalent to the grid sampling when the sampling size was large (>80).

However, model bias (mean error, ME) had shown different results among the sampling algorithms. Adaptive sampling with a moderately correlated covariate (i.e. EC_a) had a larger ME (0.003–0.008%) than grid sampling (0.003–0.006%) when the additional sampling size was small (<15), while it had a smaller ME (−0.001–0.002%) than the grid sampling (−0.002–0.003%) when the additional sampling size was intermediate to large (15–80).

Fig. 4 also shows the model performance for adaptive sampling and the simple random sampling. When a moderately correlated covariate was used (i.e. EC_a), the adaptive sampling scheme was better than the simple random sampling scheme for all the metrics (i.e. R^2, concordance, ME, RMSE).

3.4. Model performance: adaptive sampling using different priors

When different covariates (prior information) were used, the adaptive sampling schemes showed different performance. As shown in Fig. 4, when the covariate was moderately correlated with the target soil variable (i.e. EC_a), the model performance was good ($R^2 > 0.65$ and concordance >0.8 when sample size >30). However, when the covariate was weakly correlated with the target soil variable (i.e. clay content), the model performed poorly ($R^2 < 0.6$ and concordance <0.75 when sample size = 20–30), worse than the grid and simple random sampling schemes. In this case, the prior information (i.e. clay content) did not reflect the actual variation in the target soil variable (Gamma-K-ground).

It should also be noted that the effect of prior information decreased as the sampling size increased. This was not unexpected because with increasing sampling size, the regression error remained relatively constant while the kriging error decreased significantly. As such, the total error would be reduced when the additional sampling size was large enough (>80) and the LMMs performed similarly well with the grid sampling scheme.

3.5. Spatial distribution of the samples from different sampling schemes

The spatial distributions of different sampling schemes are presented in **Fig. 5**. Compared to the regular grid sampling and the simple random sampling, the adaptive sampling schemes tend to generate samples that were unevenly distributed across the field. This was because the sampling selection process of the adaptive sampling was driven by the minimizing model error within a given time interval. In general, the samples at the edge of the field were first selected to minimise the model error that is mostly attributed to kriging error. Once the samples at the edge of the field were selected, the samples in the centre were then selected. This implies that the current adaptive sampling tends to cause clustering of the samples.

The quality of the prior information had an influence on the sampling selection process. When the covariate was moderately correlated with the target soil variable (i.e. EC_a), the samples were relatively well distributed across the field. When the covariate was weakly correlated with the target soil variable (i.e. clay content), the samples were mainly located at the edge of the field. In this case, the prior information contained little useful information for the LMM and the minimization

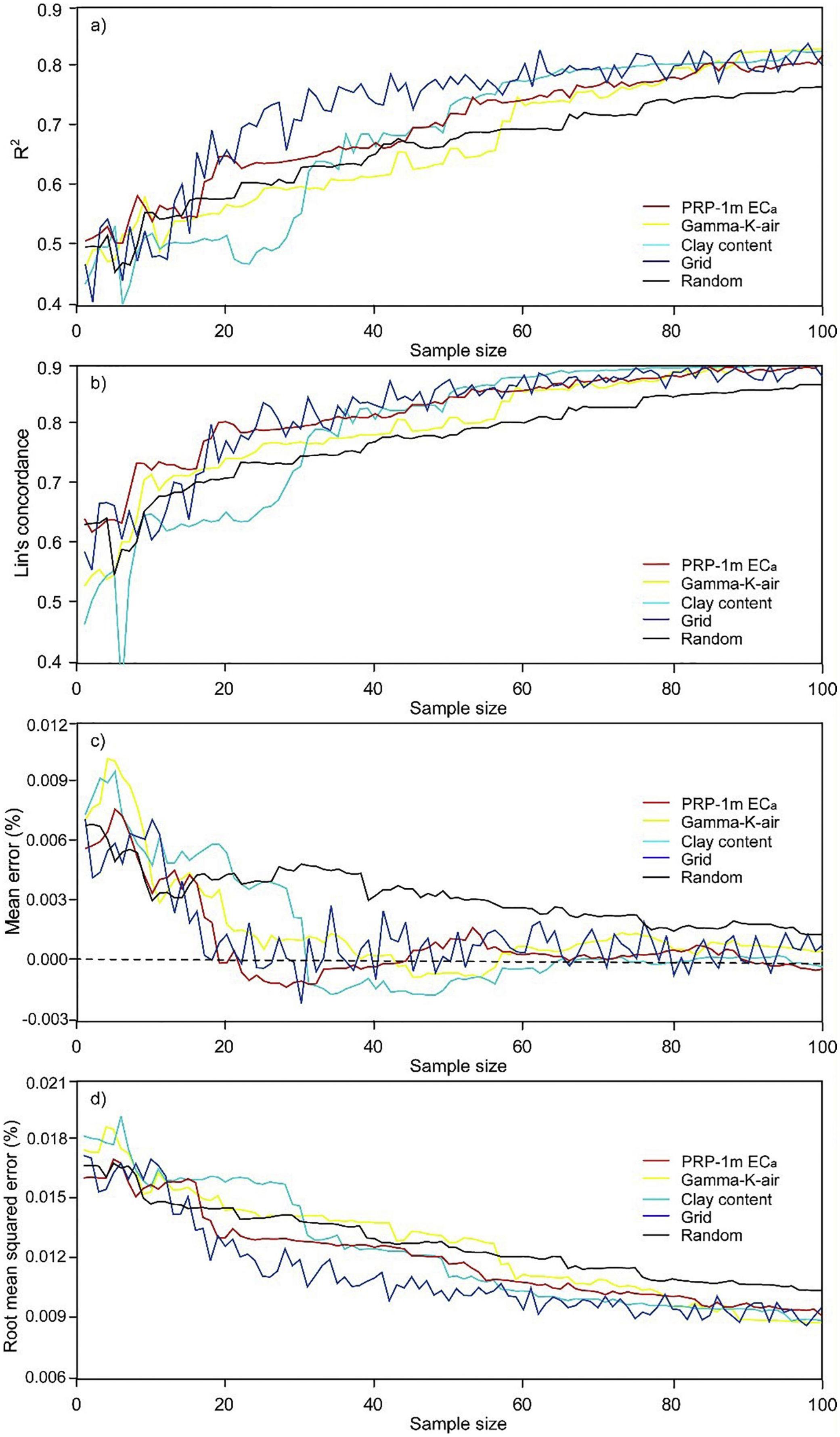

Fig. 4. Model performance of various sampling algorithms versus sampling size and including; a) R^2, b) concordance correlation coefficient, c) mean error (%), and d) root mean squared error (%). Note: Lines with different colours represent models established using different sampling algorithms, including grid and simple random sampling, adaptive sampling with covariates of different quality (e.g. PRP-1 m EC$_a$, Gamma-K-air, and clay content).

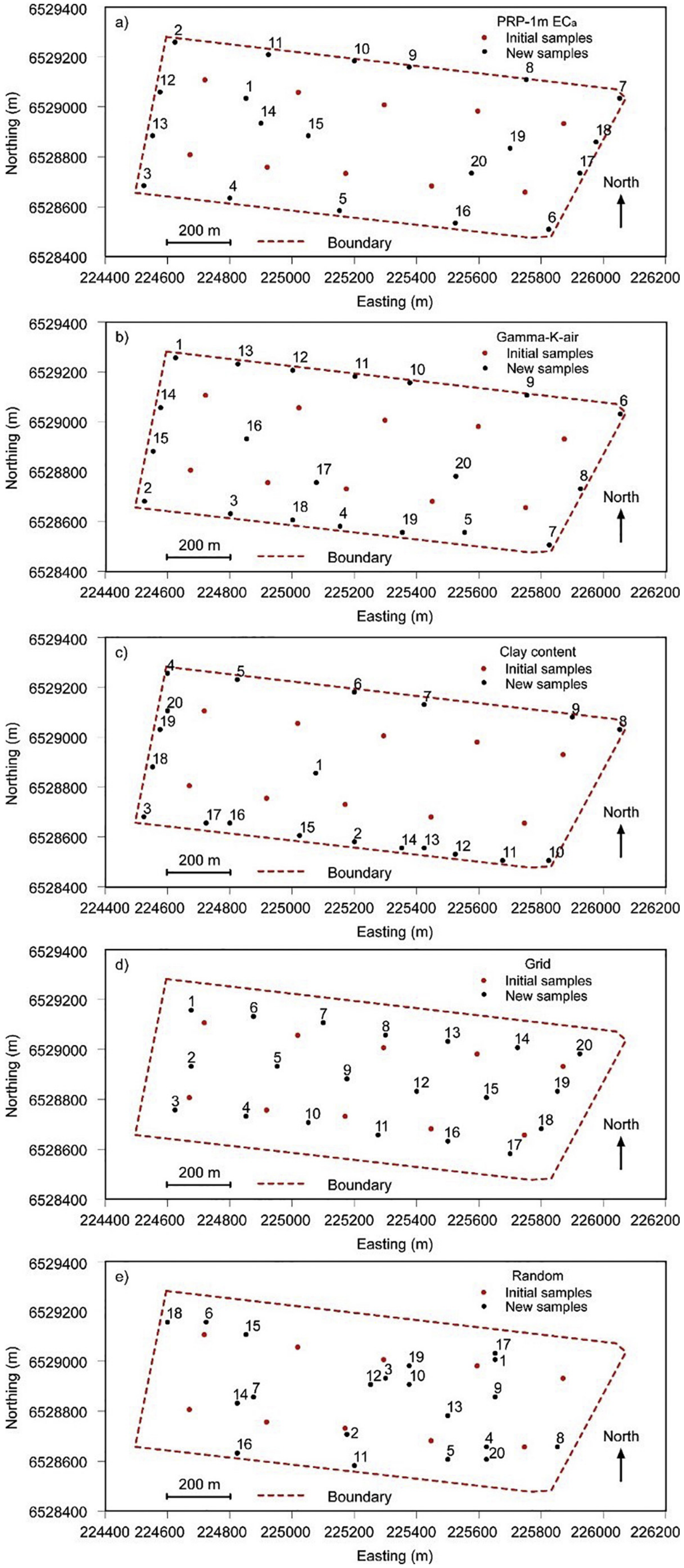

Fig. 5. Spatial distribution of the common initial 10 samples and 20 additional samples collected using different sampling algorithms and including; adaptive sampling with a) PRP-1 m EC$_a$, b) Gamma-K-air, and c) clay content, and d) grid sampling and e) simple random sampling.

Table 2
Comparison of different sampling schemes used in this study.

Sampling schemes	Sampling locations	Use of prior information	Prediction accuracy with small sampling size	Modification/addition of new samples	Travel time during sampling
Grid	Fixed	No	Low	Difficult	Short
Random	Fixed	No	Low	Difficult	Short
Adaptive	Adaptive	Yes	High	Easy	Long

of model error was mostly due to the reduction of kriging error. As such, the algorithms tend to select samples at the edge of the field to achieve a better estimation of the target variable within a target time limit.

In addition, it should be noted that the number of initial samples was set to 10. In this study, we did not investigate the number of initial samples for the configuration of the adaptive sampling scheme because the sampling locations would also vary with the quality of the covariates. It is expected that an increase of initial samples will improve the model performance (e.g. R^2) when the sample size is small (<15). Afterwards, the rate of increasing model performance with increasing sample size will become slightly smaller (approaching to the maximum model performance) for the large initial sample size as compared to that under a small initial sample size.

3.6. Advantages and disadvantages of different sampling schemes

The comparison between different sampling schemes are presented in Table 2. Compared to traditional grid-based and simple random sampling schemes, the adaptive sampling schemes used in this study have some advantages. First, the adaptive sampling makes use of the prior information of the field and generates more accurate estimation of target soil properties when the sampling size is small (<15). This suggests that the adaptive sampling scheme may reduce the sampling cost when the sampling process ("ground-truthing" or "labelling") is time-consuming and cost-prohibitive.

Secondly, the adaptive sampling scheme is adaptive. This enables subsequent sampling effort to be made when previous sampling schemes do not achieve a satisfactory model performance. This is not applicable to the traditional grid sampling scheme whereby adding new samples to the existing samples will change the whole sampling design.

There are a few disadvantages of the adaptive sampling scheme. First, the adaptive sampling scheme may generate predictive models that are biased when the sampling size is (too) small. When the travelling time is a limiting factor, the adaptive sampling scheme is worse than a regular grid sampling. This is because the adaptive sampling scheme will firstly select the points at the edges of the field represented by large kriging variance values and prioritize the points closer to the previously selected points when the travelling time becomes too long.

Secondly, the adaptive sampling scheme may expend more travelling time as compared to the grid and simple random sampling schemes. In theory, a shortest route can be calculated and implemented by a robot to survey the whole field within a time limit when the total number of samples are determined by the regular grid and simple random sampling schemes. However, this cannot be done for the adaptive sampling scheme because the location of the subsequent samples are not known a priori, it is determined from the previously collected samples and prior information. This may lead to a potential drawback of the algorithm when the travelling time is a limiting factor than the sampling cost (e.g. travelling on tough terrains or via airborne remote sensing platforms). An alternative non-adaptive sampling approach can be found in Brus and Heuvelink (2007) where simulated annealing was used to reach a balance between optimisation of the sample pattern in geographic and feature space by minimizing the spatial average (or sum) of the universal kriging variance at points. As the samples are selected to cover the geographic space, the total travelling time can be short for this non-adaptive sampling approach (Brus and Heuvelink, 2007).

Thirdly, the current adaptive sampling scheme tends to visit the edges of the field first due to the model uncertainty estimated by a kriging approach. This is not ideal and could be modified in the future by replacing the kriging approach with some machine learning algorithms that do not have an edge effect. This aspect requires further investigation. One potential solution is to use algorithms such as quantile random forest (Meinshausen, 2006; Vaysse and Lagacherie, 2017) so that model uncertainty (not affected by the location of existing samples) can be calculated across the field and used to guide the determination of the subsequent samples.

Lastly, the current adaptive sampling scheme and the linear mixed model treat all the covariates (prior information) equally regardless of their quality (correlations with the target variables). Future work can be done to use Bayesian frameworks to include the uncertainty of the prior information in the predictive models (Vrugt et al., 2009; Yang et al., 2015). In addition, non-linear regression models to account for the complex relationship between target soil variables and covariates (Archontoulis and Miguez, 2015). This would make the adaptive sampling algorithm more widely applicable.

3.7. Caveats and implications for automated agricultural management

In general, the adaptive sampling scheme can be used when the relative travelling time is short. When the sampling cost is high, the adaptive sampling scheme used in this study can produce more accurate (smaller RMSE) but more biased (larger ME) estimations of the target variables. The adaptive sampling can also be used when continuous monitoring of variables is needed whereby subsequent samples can be designed based on the previous samples to minimise the total model error, such as soil quality monitoring (Morvan et al., 2008) and census (Brawn and Robinson, 1996).

4. Conclusions

A prior-based adaptive adaptive sampling algorithm was evaluated in comparison with the grid sampling and simple random sampling for automated environmental management. We conclude:

- The adaptive sampling scheme was superior to the grid and simple random sampling schemes in terms of the accuracy of the linear mixed model when the sampling size was small (< 15 additional samples) due to the use of prior information that was well correlated with the target soil variable.
- The accuracy of the linear mixed models associated with the adaptive sampling schemes deteriorated when the correlation between the target soil variable and the prior information decreased.
- The algorithm has the potential to be applied elsewhere for automated adaptive sampling design when sampling cost is expensive and travelling time of the sensor is small.

Acknowledgments

This work was conceived at the University of Sydney. The authors acknowledge Glen D Foxwell (The University of Sydney) for his support with the field work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2020.e00284.

References

Archontoulis, S.V., Miguez, F.E., 2015. Nonlinear regression models and applications in agricultural research. Agron. J. 107 (2), 786–798.

Bartsch, E. R., Fisher, C. W., France, P. A., Kirkpatrick, J. F., Heaton, G. G., Hortel, T. C., ... & Stigall, J. R. (2002). U.S. Patent No. 6,459,955. Washington, DC: U.S. Patent and Trademark Office.

Berni, J.A.J., Zarco-Tejada, P.J., Suárez, L., González-Dugo, V., Fereres, E., 2009. Remote sensing of vegetation from UAV platforms using lightweight multispectral and thermal imaging sensors. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 38 (6), 6.

Biswas, A., Zhang, Y., 2018. Sampling designs for validating digital soil maps: a review. Pedosphere 28 (1), 1–15.

Brawn, J.D., Robinson, S.K., 1996. Source-sink population dynamics may complicate the interpretation of long-term census data. Ecology 77 (1), 3–12.

Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80 (1–2), 1–44.

Brus, D.J., Heuvelink, G.B., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1–2), 86–95.

Cohn, D.A., Ghahramani, Z., Jordan, M.I., 1996. Active learning with statistical models. J. Artif. Intell. Res. 4, 129–145.

Corwin, D.L., Scudiero, E., 2019. Mapping soil spatial variability with apparent soil electrical conductivity (ECa) directed soil sampling. Soil Sci. Soc. Am. J. 83 (1), 3–4.

Corwin, D.L., Lesch, S.M., Shouse, P.J., Soppe, R., Ayars, J.E., 2003. Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity. Agron. J. 95 (2), 352–364.

Cosh, M.H., Jackson, T.J., Bindlish, R., Prueger, J.H., 2004. Watershed scale temporal and spatial stability of soil moisture and its role in validating satellite estimates. Remote Sens. Environ. 92 (4), 427–435.

Cox, L.A., 1999. Adaptive spatial sampling of contaminated soil. Risk Anal. 19 (6), 1059–1069.

Fang, S., Da Xu, L., Zhu, Y., Ahati, J., Pei, H., Yan, J., Liu, Z., 2014. An integrated system for regional environmental monitoring and management based on internet of things. IEEE Trans. Indust. Inf. 10 (2), 1596–1605.

Flajolet, P., 1990. On adaptive sampling. Computing 43 (4), 391–400.

Grundy, M.J., Rossel, R.V., Searle, R.D., Wilson, P.L., Chen, C., Gregory, L.J., 2015. Soil and landscape grid of Australia. Soil Res. 53 (8), 835–844.

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of things (IoT): a vision, architectural elements, and future directions. Futur. Gener. Comput. Syst. 29 (7), 1645–1660.

Hart, J.K., Martinez, K., 2006. Environmental sensor networks: a revolution in the earth system science? Earth Sci. Rev. 78 (3–4), 177–191.

Jones, J. L., Mack, N. E., Nugent, D. M., & Sandin, P. E. (2005). U.S. Patent No. 6,883,201. Washington, DC: U.S. Patent and Trademark Office.

Kato, S., Tsugawa, S., Tokuda, K., Matsui, T., Fujii, H., 2002. Vehicle control algorithms for cooperative driving with automated vehicles and intervehicle communications. IEEE Trans. Intell. Transp. Syst. 3 (3), 155–161.

Kramer, H.J., 2002. Observation of the Earth and its Environment: Survey of Missions and Sensors (Springer Science & Business Media).

Kroemer, O.B., Detry, R., Piater, J., Peters, J., 2010. Combining active learning and reactive control for robot grasping. Robot. Auton. Syst. 58 (9), 1105–1116.

Kulick, J., Toussaint, M., Lang, T., Lopes, M., 2013. Active Learning for Teaching a Robot Grounded Relational Symbols (In Twenty-Third International Joint Conference on Artificial Intelligence).

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.

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.

Martinez-Cantin, R., de Freitas, N., Doucet, A., Castellanos, J.A., 2007. Active policy learning for robot planning and exploration under uncertainty. Robotics: Science and Systems 3, 321-328.

McBratney, A., Whelan, B., Ancev, T., Bouma, J., 2005. Future directions of precision agriculture. Precis. Agric. 6 (1), 7–23.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Minasny, B., McBratney, A.B., Whelan, B.M., 2006. Vesper version 1.62. Australian Centre for Precision agriculture. The University of Sydney, NSW, Australia.

Minty, B.R.S., 1997. Fundamentals of airborne gamma-ray spectrometry. AGSO J. Aust. Geol. Geophys. 17, 39–50.

Minty, B., Franklin, R., Milligan, P., Richardson, M., Wilford, J., 2009. The radiometric map of Australia. Explor. Geophys. 40 (4), 325–333.

Morvan, X., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A., ... Kibblewhite, M.G., 2008. Soil monitoring in Europe: a review of existing systems and requirements for harmonisation. Sci. Total Environ. 391 (1), 1–12.

Musafer, G.N., Thompson, M.H., 2016. Optimal adaptive sequential spatial sampling of soil using pair-copulas. Geoderma 271, 124–133.

Petersen, H., Wunderlich, T., Attia al Hagrey, S., Rabbel, W., 2012. Characterization of some middle European soil textures by gamma-spectrometry. J. Plant Nutr. Soil Sci. 175 (5), 651–660.

Pettorelli, N., Laurance, W.F., O'Brien, T.G., Wegmann, M., Nagendra, H., Turner, W., 2014. Satellite remote sensing for applied ecologists: opportunities and challenges. J. Appl. Ecol. 51 (4), 839–848.

Piikki, K., Söderström, M., Stenberg, B., 2013. Sensor data fusion for topsoil clay mapping. Geoderma 199, 106–116.

Rebeiro, P.J., Diggle, P.J., 2001. geoR: A package for geostatistical analysis. R-NEWS 1, 15–18.

Pracilio, G., Adams, M.L., Smettem, K.R., Harper, R.J., 2006. Determination of spatial distribution patterns of clay and plant available potassium contents in surface soils at the farm scale using high resolution gamma ray spectrometry. Plant Soil 282 (1–2), 67–82.

Potts, N.J., Gullikson, A.L., Curran, N.M., Dhaliwal, J.K., Leader, M.K., Rege, R.N., ... Kring, D.A., 2015. Robotic traverse and sample return strategies for a lunar farside mission to the Schrödinger basin. Adv. Space Res. 55 (4), 1241–1254.

Rahimi, M., Pon, R., Kaiser, W.J., Sukhatme, G.S., Estrin, D., Srivastava, M., 2004, April. Adaptive sampling for environmental robotics. IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 4. IEEE, pp. 3537–3544.

Salganicoff, M., Ungar, L.H., Bajcsy, R., 1996. Active learning for vision-based robot grasping. Mach. Learn. 23 (2–3), 251–278.

Schuler, U., Erbe, P., Zarei, M., Rangubpit, W., Surinkum, A., Stahr, K., Herrmann, L., 2011. A gamma-ray spectrometry approach to field separation of illuviation-type WRB reference soil groups in northern Thailand. J. Plant Nutr. Soil Sci. 174 (4), 536–544.

Searle, R., Hempel, J., Forges, A.R.D., McBratney, A.B., 2014. The Australian site data collation to support the GlobalSoilMap. In: Arrouays, D., McKenzie, N. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 127–132.

Singh, P., Gupta, A., Singh, M., 2014. Hydrological inferences from watershed analysis for water resource management using remote sensing and GIS techniques. Egypt. J. of Remote Sens. Space Sci. 17 (2), 111–121.

Spadoni, M., Voltaggio, M., 2013. Contribution of gamma ground spectrometry to the textural characterization and mapping of floodplain sediments. J. Geochem. Explor. 125, 20–33.

Tokekar, P., Vander Hook, J., Mulla, D., Isler, V., 2016. Sensor planning for a symbiotic UAV and UGV system for precision agriculture. IEEE Trans. Robot. 32 (6), 1498–1511.

Tong, S., Chang, E., 2001. Support vector machine active learning for image retrieval. Proceedings of the ninth ACM International Conference on Multimedia, pp. 107–118.

Tong, S., Koller, D., 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66.

Triantafilis, J., Gibbs, I., Earl, N., 2013. Digital soil pattern recognition in the lower Namoi valley using numerical clustering of gamma-ray spectrometry data. Geoderma 192, 407–421.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

Vrugt, J.A., Ter Braak, C.J., Gupta, H.V., Robinson, B.A., 2009. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic environ. Res. Risk Assess 23 (7), 1011–1026.

Wang, H., Zhang, T., Quan, Y., Dong, R., 2013. Research on the framework of the environmental internet of things. Int. J. Sustain. Dev. World Ecol. 20 (3), 199–204.

Webster, R., Lark, R.M., 2012. Field Sampling for Environmental Science and Management. Routledge.

Wilford, J., 1995. Airborne gamma-ray spectrometry as a tool for assessing relative landscape activity and weathering development of regolith, including soils. AGSO Res. News 22, 12–14.

Wilford, J.R., Bierwirth, P.E., Craig, M.A., 1997. Application of airborne gamma-ray spectrometry in soil/regolith mapping and applied geomorphology. AGSO J. Aust. Geol. Geophys. 17 (2), 201–216.

Wong, M.T.F., Harper, R.J., 1999. Use of on-ground gamma-ray spectrometry to measure plant-available potassium and other topsoil attributes. Soil Res. 37 (2), 267–278.

Yang, W.H., Clifford, D., Minasny, B., 2015. Mapping soil water retention curves via spatial Bayesian hierarchical models. J. Hydrol. 524, 768–779.