Geoderma 453 (2025) 117148

Contents lists available at ScienceDirect

Geoderma

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

Adding soil sampling to household surveys: Information for sample design from pilot data

R.M. Lark [a,*], L. Mlambo [a], H. Pswarayi [a], D. Zardetto [b], S. Gourlay [b]

[a] School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK
[b] Development Data Group, World Bank, Via Labicana 110, Rome, 00184, Italy

ARTICLE INFO

Handling Editor: Cristine Morgan

Dataset link: https://github.com/rmlark/Soil-Sampling-in-Household-Surveys, https://doi.org/10.6084/m9.figshare.15911973

Keywords:
Sample surveys
Soil organic carbon
Soil pH
Variance components
Sampling design

ABSTRACT

Large sample surveys with households, or individuals within households, as the basic sampled units, are important sources of information on variables related to household income, economic activity, food security and nutritional status. In many circumstances the advantages of supplementing these surveys with sampling of the soil from fields or other land units which the households cultivate may seem obvious, as a source of information on the quality of the soil on which households depend, and potential limitations on their food security such as soil pH or nutrient status. However, it is not certain that household surveys, designed to examine social and economic variables, will be efficient for collecting soil information, or will provide adequate estimates of soil property means at scales of interest. Additional sampling might be necessary, so an attendant question is whether this is feasible. In this paper we use data on soil pH and soil carbon inferred by spectral measurements on soil specimens collected from land cultivated by households in Uganda and Ethiopia to estimate variance components for these properties, and from these the standard errors for mean values at District (Uganda) or Zone (Ethiopia) level by household surveys with different designs. Similar calculations were done for direct measurement of soil carbon and soil pH from a spatial sample in Malawi from which variograms were used to infer the variance components corresponding to the levels of a household survey. The results allow the calculation of sample sizes at different levels of the design, required to allow estimates of particular quantities to be obtained with specified precision. The numbers of sampled enumeration areas required to obtain estimates of district or zone-level means with the arbitrary specified precision were large, but the feasibility of such sampling must be judged for a particular application, and the precision appropriate for that. The presented method makes that possible.

1. Introduction

Information about the soil is essential to support policy and management decisions in pursuit of the sustainable development goals (Lal et al., 2021). For example, information about the nutrient content of soils, and their capacity to retain nutrients, can support efforts to address nutrient limitations on crop production while avoiding waste and environmental impacts of unnecessary inputs. Similarly, the emergence of soil acidification or salinization can be a significant threat to soil quality, the use of land for production and the nutrient use efficiency of plants. Soil organic carbon is an important indicator of soil quality, affects the contribution of the soil to the greenhouse gas budget of agriculture, and also influences the supply of nitrogen to crops from the soil. The soil also contains micronutrients, such as zinc or selenium which are essential for human health, and soil content of these elements, and soil properties which influence their mobility, can be factors in 'hidden hunger', micronutrient deficiency alongside adequacy of dietary energy and protein (Gödecke et al., 2018).

Most food in Africa is produced on farms of about 1 ha or less (Giller et al., 2021), so there is value in soil information observed at the scale of the household and its associated fields. This is illustrated by results on field-scale measurements of soil properties and micronutrient status of the populations of Ethiopia and Malawi from Gashu et al. (2021). However, soil data are generally sparse in the global south, and while recent efforts to provide soil information through digital soil mapping (DSM) have produced the Soil Grids and iSDA information layers (Hengl et al., 2021; iSDA, 2024), these remain unvalidated. Furthermore, DSM products are essentially static, and additional sampling is needed to monitor change in the soil.

National household sample surveys are undertaken across many countries in Africa and elsewhere to provide health, demographic and

[*] Corresponding author.
E-mail address: murray.lark@nottingham.ac.uk (R.M. Lark).

https://doi.org/10.1016/j.geoderma.2024.117148
Received 15 July 2024; Received in revised form 10 December 2024; Accepted 13 December 2024
Available online 28 December 2024

social information to support development. Examples are the World Bank's Living Standards Measurement Study (LSMS) and the LSMS-ISA (Integrated Surveys on Agriculture) as described by Carletto and Gourlay (2019). If soil sampling were added to the protocols for such surveys then soil information could be collected routinely, for use in agricultural productivity analyses, for example, representing the soil under small-holder cultivation. However, the question remains whether soil samples collected in the framework of household surveys optimized to represent social and economic variables, would allow sufficiently precise estimates of soil properties, given the variability of soil, and how many households within such a survey would have to be sampled. This must be addressed systematically because of the logistical and analytical costs of soil survey can be significant, and adding components to the survey increases the respondent burden and the risk of non-participation (Singh et al., 2022).

In this study we assess the scope to estimate administrative-area means of soil properties through the integration of soil analysis in household surveys. This is done on the basis of estimated variance components for two key soil properties (pH and soil organic carbon: SOC) in topsoil and subsoil. Data were used from two experimental surveys in which soil sampling was carried out, within a household survey sampling frame, on land owned by households. We also assessed the potential of DSM products to assess soil variation for purposes of planning soil sampling as part of a household survey, and used a geostatistical model of soil variation in Malawi, from data collected in a spatial coverage survey not aligned with household surveys, to estimate the variance components required for this sampling planning task. These results allowed us to assess the expected precision of soil information based on sampling in a household survey, and to identify the intensity of surveying required in such a framework to provide adequate information at the target scale (administrative area).

2. Methods

2.1. Data from household-aligned surveys

The data from household surveys were from two aligned projects in Uganda and Ethiopia. These were two-stage sample surveys in which primary sample units (enumeration areas: EA) were selected from a sample frame comprising all EA in the administrative unit. The secondary sample units (households: HH) were then selected from among HH within the selected primary units. Details of the designs are in Sections 2.1.1 and 2.1.2. These studies were designed specifically to evaluate different methods to measure key properties of agricultural production systems, including soil health, and so the sampling densities (in terms of the number of EA per administrative unit) were not typical of national household surveys, with more sample effort than would be usual per unit area. For example, in the methodological studies 15–29 EA were sampled per district whereas in the Uganda National Panel Survey (UNPS) 2–3 EA are typically sampled per district, and 2–6 in the Uganda Household Integrated Survey (UHIS). Similarly, in Ethiopia, typically 2–3 EA are selected per Zone in the Ethiopia Socioeconomic Survey, with somewhat more intensive sampling in the Ethiopian Agricultural Sample Survey (Ag-SS) at 6–24 EA per Zone. This relatively dense sampling in the methodological studies is an advantage, because the more intensive sampling provides data from which to estimate variance components at between-EA and between-HH within-EA levels with confidence. These variance components can then be used to explore the expected precision of estimates based on surveys of differing intensity.

2.1.1. MAPS 2015 survey and data, Uganda

The MAPS 2015 project (Methodological Experiment on Measuring Maize Productivity, Variety and Soil Fertility) was undertaken in Uganda as a partnership between the World Bank LSMS-ISA programme and the Uganda Bureau of Statistics. More information about the survey is provided by Gourlay et al. (2019). The 2015 data used here were collected from three units in eastern Uganda: Serere district, Sironko district and one unit comprising part of Iganga and Mayuge districts. A total of 75 enumeration areas (EA) were selected by random sampling with inclusion probability proportional to size (PPS) giving 15 EA in each of Serere and Sironko districts and 45 in Iganga–Mayuge. Within each EA twelve households (HH) were selected independently and at random from a listing with the objective of selecting six from among those with monocrop maize plots and six from those with intercropped maize plots. One field was selected at random from each household, from among its monocrop or intercrop plots depending on the subset to which it belonged.

Soil samples were collected from the selected plot for each household following crop planting unless heavy rains made this impossible in which case soil was collected when the crop was cut for yield estimation. Soil was sampled at a central location in the plot at depths 0–20 cm and 20–50 cm. An additional three samples from 0–20 cm were collected at locations 12.2 m from the central location, the first upslope from the centre, and the second and third on bearings 120 degrees and 240 degrees from the upslope direction. The four samples from 0–20 cm were bulked in the field, then coned and quartered to a subsample of approximately 120 g. The soil samples were double-bagged and barcoded, then delivered to the analytical laboratory within five days of collection.

The final data set, after removing observations missing a unique HH designation, comprised data from 892 households. Of these 874 had a topsoil (0–20 cm) sample and a subsoil (20–50 cm) sample, 8 had a topsoil sample only and 10 had a subsoil sample only.

2.1.2. LASER survey and data, Ethiopia

The Ethiopia Land and Soil Experimental Research (LASER) study was carried out by the Central Statistical Agency of Ethiopia (CSA) in 2013 in partnership with the World Bank's LSMS team. This was a pilot study primarily focused on methods to estimate maize yield and productivity. Information about the survey is provided by Central Statistical Agency (2017).

The LASER project was conducted in Oromia Region of Ethiopia over three administrative Zones: Borena, West Arsi and East Wellega. A total of 85 enumeration Areas (EA) were chosen from these Zones from the universe of EA that were included as part of the Agricultural Sample Survey, in which EA were selected with probability proportional to size (Central Statistical Agency, 2017). Twelve HH were selected in each EA and from each HH up to two fields were selected. One field was selected at random from among the HH's fields under monocrop maize, and a second field was then selected at random from all other fields cultivated by the HH. It was possible that a HH cultivated fields in more than one distinct parcel of land, but the sampling was not structured by parcels, and so observations from different parcels within a household arose at random and not from the sample design.

Exploration of the data set found that for each of 297 HH just one field had been sampled, and for each of 688 HH two fields had been sampled. In each of 251 of the HH two fields had been sampled from different parcels of land, and for the remaining 437 the samples were from two fields within the same parcel of land.

2.1.3. Soil data and their exploratory analysis

The soil samples collected from both the LASER and MAPS surveys were analysed for different properties. Here we focus on two key soil variables which are indicators of quality and potential limitations on crop production: soil pH and organic carbon content. The values for these variables for each soil sample were predicted at the ICRAF Soil-Plant Diagnostic Laboratory. Protocols for that Laboratory's procedures are provided by Ateku (2021a,b) and Ateku and Chacha (2021). Near- and mid-infrared spectral measurements were made on each sample, and a subset of ten percent were also analysed for the properties by conventional wet chemistry to provide values to calibrate the spectral data. Soil pH was determined for the calibration subset in a 1:2 water suspension, and soil organic carbon was determined after acidification to remove inorganic carbon in the form of carbonate. All the data used from these surveys in the rest of the paper are the predicted values based on the spectral measurements.

Summary statistics on the data were computed. In particular the conventional skewness coefficient was estimated as a measure of the asymmetry of the data distribution. Conventionally, steps such as transformation are considered if this falls outside the interval $[-1, 1]$. However, as the skewness coefficient may be susceptible to the effect of outlying observations we also computed the octile skewness of Brys et al. (2003). Transformation may be required when this falls outside the range $[-0.2, 0.2]$ (Rawlins et al., 2005).

Probable outliers in each data set, in the sense of Tukey (1977), were found as observations outwith the 'outer fences' of the data. The outer fences are at

$$Q1 - 3H,\ Q3 + 3H,$$

where $H = Q3 - Q1$ and $Q1, Q3$ are the first and third quartiles of the data respectively.

A histogram and boxplot were produced for each variable, in addition to a QQ plot: a plot of the empirical quantile values represented by each standardized datum against the equivalent quantile of a standard normal distribution. Data from a normal random variable should fall on a straight line.

These plots and summary statistics were used to decide whether the data could plausibly be regarded as a normal random variable on the original scales of measurement, or whether a transformation to normality was required prior to further analysis.

2.1.4. Linear mixed models: model structure and estimation

In both these projects, Districts/Zones were selected for sampling and then EA within the districts, HHs within EA and Fields within HH were selected by random sampling. This can be represented in a linear mixed model (LMM) with an overall mean as a fixed effect, and the District/Zone, EA, HH and field-level variation represented by random effects. In these studies we also included the crop management system (monocrop or intercrop) as a fixed effect.

The LMM for a vector of n observations, $\mathbf{y}$, takes the following form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is an $n \times p$ design matrix which associates each of the n observations with a value for each of p fixed effects, and $\boldsymbol{\tau}$ contains the fixed effects coefficients. In this case the fixed effects are an intercept, the mean value of the soil property under the reference management (monocrop or intercrop), and an additional fixed effect coefficient which is the additive effect of the non-reference management.

There are r random effects, with values in $\mathbf{u}$, and $\mathbf{Z}$ is an $n{\times}r$ design matrix which associates each observation with a subset of these. In the case of the MAPS survey the random effects are the r_D districts, r_EA EAs and the r_HH HHs, with EA nested within the Districts and HH nested within EA. We may therefore think of $\mathbf{u}$ as comprising three subvectors, the $r_\mathrm{D} \times 1$ vector of District random effects, $\mathbf{u}_\mathrm{D}$, the $r_\mathrm{EA} \times 1$ vector of EA random effects, $\mathbf{u}_\mathrm{EA}$, and the $r_\mathrm{HH} \times 1$ vector of HH random effects, $\mathbf{u}_\mathrm{HH}$ where $r = r_\mathrm{D} + r_\mathrm{EA} + r_\mathrm{HH}$ and $\mathbf{u} = \left[\mathbf{u}_\mathrm{D}{}^\mathrm{T}, \mathbf{u}_\mathrm{EA}{}^\mathrm{T}, \mathbf{u}_\mathrm{HH}{}^\mathrm{T}\right]^\mathrm{T}$. The design matrix $\mathbf{Z}$ associates each observation with exactly one District, one EA within that District and one HH within that EA. The term $\boldsymbol{\varepsilon}$ is an independent and identically distributed residual, this represents the between-field within HH variation of the observations, including analytical error.

In the case of the MAPS survey there were a significant number of households where the sampled fields were selected from distinct parcels, and so an additional between-parcel within HH random effect can be specified. However, for purposes of simplicity in presentation we stick with the case with just three random effects at District, EA within-District and HH within-EA level.

It is assumed that the random effects and residual terms are independent, and normally distributed. The random effects at each level of a nested structure have a common variance, so in this case the model parameters include a separate variance component for Districts (Zones in the LASER survey) σ^2_D; one for EA within Districts, $\sigma^2_\mathrm{EA:D}$; one for HH within EA $\sigma^2_\mathrm{HH:EA}$ and a residual variance σ^2_ε. The joint distribution of $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ is therefore modelled as

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} \sim \mathcal{N}\left\{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_\upsilon\end{bmatrix}\right\}, \tag{1}$$

where the random effects have an $r \times r$ covariance matrix $\mathbf{G}$. In the case of the nested design-based sample used here G can be written in terms of the variance components for the random effects and identity matrices $\mathbf{I}$:

$$\mathbf{G} = \begin{bmatrix}\sigma^2_\mathrm{D}\mathbf{I}_{r_\mathrm{D}} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma^2_\mathrm{EA:D}\mathbf{I}_{r_\mathrm{EA}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \sigma^2_\mathrm{HH:EA}\mathbf{I}_{r\mathrm{HH}}\end{bmatrix},$$

and the residual term has an $n{\times}n$ covariance matrix $\mathbf{R}_\upsilon = \sigma^2_\varepsilon\mathbf{I}_n$, which is diagonal because of the assumption that the residuals are independent. The unknown parameters, the variance components σ^2_EA, $\sigma^2_\mathrm{HH:EA}$ and σ^2_ε are estimated by residual maximum likelihood which avoids the well-known bias in ordinary maximum likelihood estimation. We did this analysis using the lme function from the nlme library for the R platform (Pinheiro and Bates, 2000).

2.1.5. Survey weights

The original weights from these surveys were not available. This does not prevent our analysis of the data by the LMM in which, rather than using sample weights derived from the inclusion probabilities, we use modelled variance components as a basis for weighting the contribution of observations to estimates. The variance components for the LMM model, estimated as described above, provide a valid basis for quantifying the precision of model-based estimates and predictions from new data collected according to some comparable hierarchical design and assumed to be a realization of the same model.

Sample weights can be incorporated into the fitting of a linear mixed model to data from a design-based survey (see Carle, 2009). To do this the weights are rescaled, for which there are two principal methods. If $w_{i,j}$ is the sample weight for observation i in cluster j then the two rescaling schemes are as follows:

$$w^\mathrm{A}_{i,j} = w_{i,j}\left(\frac{n_j}{\sum_i w_{i,j}}\right), \tag{2}$$

$$w^\mathrm{B}_{i,j} = w_{i,j}\left(\frac{\sum_i w_{i,j}}{\sum_i w^2_{i,j}}\right). \tag{3}$$

Neither of these weighting schemes is regarded as generally best. Carle (2009) suggests that the weights $w^\mathrm{A}_{i,j}$ are preferable for point estimates, whereas the weights $w^\mathrm{B}_{i,j}$ will provide better estimates of variance components. In this study we used both weighting schemes for comparison with outputs from an unweighted fitting of the model. The LMM with weights were obtained using the WeMix library in R (Bailey et al., 2023; R Core Team, 2023). This can be used for LMM with random effects at no more than three levels, so for the LASER data from Ethiopia, the Household and Parcels had to be absorbed into the residual for analysis in WeMix.

In lieu of survey weights, which were not available for the MAPS and LASER studies, approximate sampling weights for the probability of EA selection were estimated. Note that these estimated probabilities of selection, which are based on a number of assumptions, only go so far as the EA selection, and do not incorporate household or plot level selection probabilities, which were assumed to be uniform within each EA. For LASER, we backed out the estimated probability of inclusion of a given EA in the AgSS survey which was sampled PPS (with the assumption that all EA within a given woreda have the same population), and multiplied that by the probability of selection into the LASER study given the AgSS sample and the practical allocation of EA across agroecological zones identified in the sample design. For the Uganda MAPS study, in which EA were selected with PPS, the estimated probabilities of EA selection are employing two very strong assumptions: (i) assuming one EA equals one village, and (ii) assuming all villages within a given sub-county have the same population. These are to be taken as rough approximations in assessing the implications of sample clustering within districts/zones.

2.1.6. E-BLUP and variance

Once the random effects parameters are estimated, the Mixed Model equation of Henderson et al. (1959) can be applied to obtain the Best Linear Unbiased Estimates of the fixed effects coefficients (BLUE, $\widehat{\boldsymbol{\tau}}$) and the Best Linear Unbiased Predictions of the random effects (BLUP, $\widetilde{\mathbf{u}}$). The equation is as follows:

$$\begin{bmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{Z}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\tau}}\\ \widetilde{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{y}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{R}_{v}^{-1}\mathbf{y}\end{bmatrix}. \tag{4}$$

The error covariance matrix of the estimates/predictions $\begin{bmatrix}\widehat{\boldsymbol{\tau}}^{\mathrm{T}} & \widetilde{\mathbf{u}}^{\mathrm{T}}\end{bmatrix}$, $\mathbf{C}$ is estimated by

$$\widehat{\mathbf{C}}=\left(\mathbf{W}^{\mathrm{T}}\mathbf{R}_{v}{}^{-1}\mathbf{W}+\mathbf{G}^{*}\right)^{-1}, \tag{5}$$

where $\mathbf{W}\equiv[\mathbf{X},\mathbf{Z}]$ and $\mathbf{G}^{*}\equiv\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{bmatrix}$.

The BLUP for some random quantity, described by a LMM, is the mean for the prediction distribution of that quantity, conditional on the model and observations. When the REML estimates of the random effects parameters are used to specify the model then BLUP is sometimes called the empirical BLUP or EBLUP. Eq. (4) can be solved to find the EBLUP of the individual random effects by using the estimated variance parameters to specify $\mathbf{R}_{v}$ and $\mathbf{G}$.

In this study we specify the prediction error variance of the E-BLUP of the mean for a sampled District/Zone as a quality measure on which to compare different sampling schemes. Others could be considered. In our models the District/Zone is a random effect, so we assume that the fixed effect is an overall mean and that the E-BLUP for the mean of a sampled District/Zone is therefore the combination of the BLUE of the overall mean and the BLUP for the District/Zone random effect (which is not nested in any other RE). If this sum can be computed by the operation $\lambda^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}$, where λ is a vector length $p+r$ with value 1 for the fixed effect and random effect values corresponding to the overall mean and the random effect of interest, then the variance of the error is

$$\lambda^{\mathrm{T}}\mathbf{C}^{-1}\lambda. \tag{6}$$

We computed the prediction error variance of the E-BLUP for the mean of a sampled District/Zone for the observed soil properties in the MAPS and LASER surveys, assuming the variance components estimated from the data. The sampling schemes were specified as follows. For Uganda (MAPS) we assumed a sample campaign over two districts. Four fixed sample sizes were specified. All possible combinations of the number of EA sampled per district and number of HH sampled per EA consistent with the specified sample size were considered and the standard error for the E-BLUP of the mean of a sampled district was computed. This objective function was then plotted against the number of sampled EA.

Soil organic carbon had been transformed to logarithms for analysis. Assuming the District mean to be the same as the sample mean, the interval of the mean ±1 standard error on the log scale was computed and back-transformed to the original scale of units. The width of this interval on the original scale (percent organic carbon) was plotted against the number of EA

Similarly for the LASER data from Ethiopia, the prediction error variance of the E-BLUP for a Zone mean was computed from the estimated variance components. It was assumed that two Zones were sampled, and four fixed total sample sizes were specified. Two land parcels were sampled per HH, with one field sampled in each parcel. All combinations of the number of sampled EA per Zone and HH per EA consistent with the total sample size were considered, and, for each soil property, the SE of the E-BLUP of the Zone mean was computed and plotted against the number of EA.

2.2. Estimates of variance components from alternative data sources

In the case of the MAPS 2015 survey from Uganda and the LASER survey from Ethiopia we have data from a household survey which allow us to estimate directly the variance components for a model of the soil variable which we require in order to compute prediction error variances for E-BLUPs from different sample designs. However, it is not always the case that such pilot information will be available. Is it possible to arrive at estimates of these variance components from sources of information other than household surveys? Two approaches were considered. The first was based on the use of digital soil mapping products. The second entailed the use of a geostatistical model for a target soil property derived from data collected on a sampling design selected for geostatistical mapping which does not allow direct estimation of the variance components for a household survey.

2.2.1. Digital soil information products

Digital soil maps are predictions of soil properties, including soil classes, made by some quantitative model or algorithm from a limited number of direct observations and associated covariate data such as remote sensor measurements or variables extracted from a digital terrain model. Two such digital soil information products for Africa are SoilGrids (version 2.0) with predictions on 250-m pixels (Poggio et al., 2021) and the iSDA (Hengl et al., 2021) with predictions on 30-m pixels.

It should be noted that, while considerable effort has been put into producing these DSM products, they are not widely and independently validated, and so their usefulness as predictions of soil properties are not yet established. Furthermore, any DSM product comprises predictions, based on models or algorithms, which are subject to shrinkage or smoothing and so are not expected to exhibit spatial variation directly comparable to that of the corresponding soil properties. Data on soil pH and organic carbon were extracted from the iSDA and SoilGrids2.0 digital soil map products, based on the georeferenced coordinates of each agricultural plot in which ground-based soil samples were collected.

The MAPS and LASER soil samples and corresponding analyses are for depth intervals 0–20 cm and 20–50 cm (topsoil and subsoil). The iSDA predictions are for these same depths, the SoilGrids predictions are for 0–5 cm, 5–15 cm and 15–30 cm. Topsoil values from SoilGrids, corresponding to the 0–20 cm sample interval were computed as weighted averages of the SoilGrids predictions at the three depths, with weights 0.25, 0.5 and 0.25 respectively.

The extracted predictions were analysed with a LMM as for the direct soil measurements, and the estimated variance components were used to compute the prediction standard errors of E-BLUPs of the mean for sampled Zones or Districts. These could then be compared with the standard errors based on variance components estimated from direct observations of the soil properties.

2.2.2. Point soil data from spatial surveys

Soil sampling undertaken to support spatial prediction of soil properties is not, in general, collected according to hierarchical sampling designs as these are not efficient for this objective. Rather, spatial coverage samples are preferred, to limit the mean distance from an unsampled location to the nearest observation in the sample. Such a sample may be on a regular grid, or generated by some algorithm such as the k-means algorithm in the spcosa library for the R platform (Walvoort et al., 2010). Some supplementation of a spatial coverage survey with closely-paired observations is recommended to support the estimation of a spatial statistical model of the data (Lark and Marchant, 2018). In this study we propose that data from such surveys, and other point data sets not obtained in household surveys, could be used to estimate variance components corresponding to hierarchical levels of a household survey in order to evaluate the precision of E-BLUPs based on household surveys of different size and composition. This is through the estimation of a variogram model for the target variable from the spatial data.

In this study we used point soil data from the GeoNutrition survey of soil and crops in Malawi and Ethiopia (Kumssa et al., 2022). The survey of Malawi was a spatial coverage survey across the country (the sample frame was those areas in agricultural use). Spatial coverage points were obtained with the k-means algorithm in the spcosa library referred to above, supplemented with some close-pair observations. Full details of the sampling design and its implementation are given by Kumssa et al. (2022). The survey of Ethiopia was done differently because the sample frame is more spatially fragmented than in Malawi due to the marked relief over much of Ethiopia. Rather, sample locations were selected by a random sampling design with equal inclusion probabilities for all nodes in the sample frame, but the sites selected to achieve spatial balance with spread (Grafström and Schelin, 2014). Random locations were then selected for supplementary sampling with a closely-paired location such that about 10% of the total sample comprised close-pair points.

Both the Malawi and Ethiopia GeoNutrition data sets contain point observations of soil pH (measured in water) and soil organic carbon content for depth interval 0–15 cm. These data were used to estimate variograms. A subset of data were obtained from the Ethiopia set within part of Oromia Region delineated by the minimum and maximum longitudes and latitudes for West Arsi and East Wellega respectively. This gave 122 data for each variable. A variogram was estimated for each variable by residual maximum likelihood (REML) using the likfit function from the geoR library for the R platform (Diggle and Ribeiro, 2007). This variogram function was cross-validated using the xvalid function from the same library.

In the case of Malawi, all 1812 data from across the country were used. Because the data set was large the variogram was estimated by Matheron's estimator (Matheron, 1962) and a model was fitted by weighted least squares (Webster and Oliver, 2007) and cross-validated (Lark, 2003). The estimation of variance components from the variogram of a soil property was then based on Krige's relation, as described below.

Krige's relation concerns the relationship between the dispersion variances of a variable, measured on different *supports* within some region $\mathcal{R}$. By support we denote the size and shape of a spatial unit over which the variable is (linearly) aggregated. Fig. 1 shows a region $\mathcal{R}$ and domains on one of two supports, $\mathcal{B}$ and $\mathcal{B}'$ over which a variable may be aggregated (domain means). Note that support $\mathcal{B}'$ is smaller than $\mathcal{B}$ in the sense that we could represent some domain, $\mathcal{B}_i$, on support $\mathcal{B}$, as the union of some set of domains on support $\mathcal{B}'$

$$\mathcal{B}_i \equiv \cup \left\{\mathcal{B}'_1, \mathcal{B}'_2, \dots\right\}.$$

We denote the variance of a variable, Z, on support $\mathcal{B}$ within region $\mathcal{R}$ by $D(\mathcal{B},\mathcal{R})$. In geostatistics this is called the dispersion variance (Journel and Huijbregts, 1978). Krige's relation states the additivity of dispersion variances on nested supports,

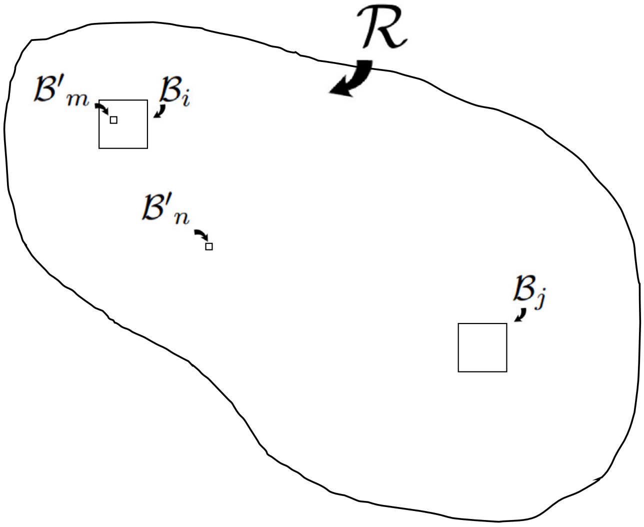

Fig. 1. A hypothetical region, $\mathcal{R}$, which includes two larger subregions $\mathcal{B}_i$ and $\mathcal{B}_j$ and smaller subregions $\mathcal{B}'_m$ and $\mathcal{B}'_n$, with $\mathcal{B}'_m$ nested in $\mathcal{B}_i$.

$$D(\mathcal{B}',\mathcal{R}) = D(\mathcal{B}',\mathcal{B}) + D(\mathcal{B},\mathcal{R}), \tag{7}$$

(Journel and Huijbregts, 1978). That is to say, the variance of Z on support $\mathcal{B}'$ in $\mathcal{R}$ can be partitioned into the variance on a larger support $\mathcal{B}$ within $\mathcal{R}$ and the variance on support $\mathcal{B}'$ within $\mathcal{B}$. The two terms on the right hand side can be thought of as variance components, so we can obtain the between $\mathcal{B}'$ within $\mathcal{B}$ variance component by

$$D(\mathcal{B}',\mathcal{B}) = D(\mathcal{B}',\mathcal{R}) - D(\mathcal{B},\mathcal{R}). \tag{8}$$

If the variogram of Z on a point support, that is to say on a support much smaller than the smallest nested domain of interest (e.g. a core within a field), is $\gamma(|\mathbf{h}|)$ then the dispersion variance (core support) within region $\mathcal{R}$, $D^{\cdot}(\mathcal{R})$, can be obtained by the double integral

$$D^{\cdot}(\mathcal{R}) = \int_{\mathbf{x}_1\in\mathcal{R}} \int_{\mathbf{x}_2\in\mathcal{R}} \gamma\left(|\mathbf{x}_1 - \mathbf{x}_2|\right) \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{x}_1, \tag{9}$$

see, for example, Webster and Oliver (2007).

From Krige's relation we can write

$$D^{\cdot}(\mathcal{R}) = D^{\cdot}(\mathcal{B}) + D(\mathcal{B},\mathcal{R}), \tag{10}$$

which can be rearranged and substituted into Eq. (8) to give an expression for the nested variance component

$$D(\mathcal{B}',\mathcal{B}) = D^{\cdot}(\mathcal{B}) - D^{\cdot}(\mathcal{B}'). \tag{11}$$

In the case of the Ethiopia dispersion variances were calculated by computing the mean value of the variogram over all pairs of points in the HH survey. These pairs could be sorted into pairs within the same EA or within different EA within the same Zone, or within different Zones, and so used to compute the dispersion variances from which the variance components corresponding to the levels of the HH survey could be estimated.

In the case of Malawi shapefiles were used for the Enumeration Areas and Districts so that $D^{\cdot}(\mathcal{R})$ could be computed across the cultivated domain of the country, and then dispersion variances within Districts and EA by Monte Carlo integration. Variance components were then used to obtain the prediction error variances of the Zone or District E-BLUP, as with the HH surveys.

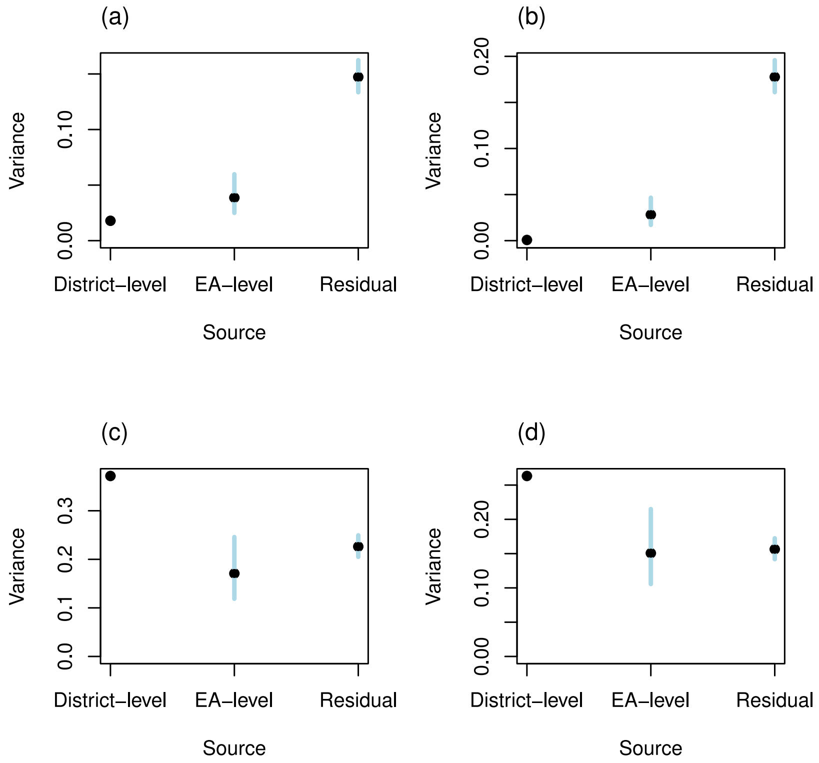

Fig. 2. Uganda data: variance components for (a) topsoil pH (b) subsoil pH (c) log topsoil SOC and (d) log subsoil SOC with 95% confidence intervals.

3. Results

3.1. Exploratory data analysis

3.1.1. Uganda

Summary statistics are presented in Table A1 in the supplementary material, with plots in Figure A1. These suggest that soil pH at both depths can plausibly be regarded as normally distributed, with some outliers in the upper tail. The small octile skewness (comfortably in $[-0.2, 0.2]$) is indicative of a symmetrical distribution, the conventional skewness is relatively large, presumably due to its susceptibility to outliers. Outliers were identified at both depths according to the criterion of Tukey (1977). Because this study is focussed on the size of variance components, which are very susceptible to outliers, these outlying data were removed before analysis. There are clear relationships between topsoil and subsoil values of soil pH (Figure A4). There is little evidence for a difference in soil pH between the monocrop and intercrop fields (Figure A6).

The exploratory plots (Figure A2) and summary statistics (Table A1) indicate that SOC is markedly skewed in the topsoil and subsoil (octile skewness of 0.38 and 0.47 respectively). This is reduced by transformation to natural log (see Table A1 and Figure A3), but there may be two distinct subpopulations, as both histograms for SOC show a bimodal distribution on the log scale. The summary statistics and the scatterplot (Figure A5) show that SOC concentration is larger in general in the topsoil than the subsoil, although there is a correlation between SOC at the two depths. The boxplots in Figure A6 indicate that SOC may be slightly larger under intercropping than under monocrop maize.

3.1.2. Ethiopia

Summary statistics (Table A2) and the exploratory plots (Figure A7) indicate that soil pH at both depths can plausibly be treated as normally distributed. The pH values at the two depths are strongly related (Figure A9), and there is some evidence for a possible difference in pH between crop management systems, with somewhat smaller pH under the monocrop maize. There is evidence of a trend in pH from south to north (Figure A12), possibly reflecting the effect on soil acidity of greater rainfall in the north.

The octile skewness for SOC in Ethiopia is small at both depths, although the conventional skewness is large (0.75 and 1.10 for the top and subsoil). There are outlying data on SOC by the criterion of Tukey (1977), and the exploratory plot (Figure A8) shows that most of the data show a symmetrical distribution, which could plausibly be regarded as normal, but with some very distinct outliers. As with the data on soil pH from Uganda, these outlying data were removed before further analysis. There was no evident effect of crop management on SOC (Figure A11), nor of a trend with latitude (Figure A12).

3.2. Linear mixed models: effects of weights, parameter estimation and inference

3.2.1. Effects of including or ignoring sample weights

The variance components for fitted LMM are presented in Table 1, including those estimated ignoring the weights and those estimated with WeMix using the approximate weights, under both schemes for rescaling the weights. Very small effects are seen from ignoring the weights for variance components at EA scale and finer. This suggests that the sample weights are not informative about the target variables. Given that we are using the prediction error variance of the E-BLUP of the mean at Zone or District level to assess the quality of a sampling design, and our uncertainty about the true weights, further analysis was based on the unweighted output. Household and Parcel were therefore included for the LASER survey analyses.

3.2.2. Uganda

The LMMs for soil pH included cropping system as a fixed effect, because of its use in the sampling design. There was no evidence to reject the null hypothesis of no difference in the soil pH under the contrasting cropping systems for either the topsoil or the subsoil ($p = 0.94$ and 0.51 respectively). Nonetheless, we retain cropping system in the model, because it cannot be regarded as a random effect. The variance components for soil pH (Table 1, Fig. 2(a,b)) were small at

Table 1
Estimated variance components with unweighted and weighted fitting of the linear mixed model.

Level	MAPS survey, Uganda					
	pH, topsoil			pH, subsoil		
	Unweighted	Weighted	Weighted	Unweighted	Weighted	Weighted
		w^A	w^B		w^A	w^B
District	0.018	0.012	0.013	0.001	<0.001	<0.001
EA	0.039	0.036	0.033	0.028	0.026	0.023
Residual	0.147	0.144	0.148	0.178	0.172	0.176
Level	log SOC, topsoil			log SOC, subsoil		
	Unweighted	Weighted w^A	Weighted w^B	Unweighted	Weighted w^A	Weighted w^B
District	0.112	0.081	0.082	0.105	0.067	0.077
EA	0.063	0.057	0.054	0.073	0.075	0.064
Residual	0.124	0.120	0.123	0.115	0.114	0.116
Level	LASER survey, Ethiopia					
	pH, topsoil			pH, subsoil		
	Unweighted	Weighted w^A	Weighted w^B	Unweighted	Weighted w^A	Weighted w^B
Zone	0.250	0.165	0.165	0.257	0.179	0.169
EA	0.159	0.160	0.159	0.213	0.213	0.212
Residual	0.302	0.300	0.297	0.290	0.288	0.288
Level	SOC, topsoil			SOC, subsoil		
	Unweighted	Weighted w^A	Weighted w^B	Unweighted	Weighted w^A	Weighted w^B
Zone	0.272	0.175	0.175	0.112	0.071	0.071
EA	0.667	0.667	0.666	0.489	0.489	0.487
Residual	0.422	0.417	0.417	0.349	0.346	0.346

district level, and only slightly larger at EA-level. The largest variance component, at both depths, was the residual (between HH). This short-range variation may reflect management practices, as well as analytical error. It is worth noting that the same pattern is seen at both depths, so there may be underlying variations in the soil material as well.

As for pH, there was no evidence to reject the null hypothesis of no difference between cropping system with respect to SOC at either depth ($p = 0.32$ and 0.55 respectively). In contrast to soil pH, the largest variance component for SOC at both depths was the between-District component. The between-EA and between-HH (residual) variance components were similar to each other at both depths.

3.2.3. *Ethiopia*

There was very weak evidence that the mean pH for topsoil under monocropped maize was slightly more acid than under intercrop ($p = 0.06$), and no evidence to reject the null hypothesis of no difference in the case of the subsoil ($p = 0.73$). The variance components for pH (Fig. 3(a,b)) from the Ethiopia survey show a similar pattern for topsoil and subsoil with the largest component at the between-zone level, large components at the between-EA and residual (between-field within parcel), and rather small variance components at HH and parcel level.

There was moderate evidence for a small difference in SOC concentration in topsoil between the soils under monocrop and intercrop maize (mean difference of 0.11%), ($p = 0.01$), but no evidence for a difference in the subsoil ($p = 0.77$). The largest variance component for SOC at both depths was at the between-EA level, and the next largest was the residual at both depths. The between-zone variance was similar to the residual in size for the topsoil, but smaller for the subsoil. As for pH, at both depths, the variance components at HH and parcel level were the smallest.

3.3. *Error of BLUPs*

3.3.1. *Uganda, district BLUPs*

For soil pH in the Ugandan setting we consider the standard error of the E-BLUP of a District mean in surveys of two Districts with different total sample sizes distributed over different numbers of EA per District (Fig. 4(a,b)). We consider a target standard error for the prediction of 0.05, so that the 95% confidence interval is about ±0.1 pH units (dashed line on the figures).

With a total sample size of 600 or fewer, this target is achieved for topsoil pH only with 20 or more EA sampled per district. With a total sample size of 100 this target is not achieved (even with 50 EA per district). With a total sample size of 200 the target is achieved with 50 or more EA per district In practice the sample intensity for HH surveys in Uganda ranges from 2–3 EA per district (Uganda National Panel Survey) and 2–6 EA per district (Uganda Household Integrated Survey). At these intensities the standard error for the BLUP of topsoil pH varies from 0.08 to 0.13, giving wider confidence intervals than the target.

Variation of soil pH in the subsoil was less pronounced than in the topsoil and so the target standard error can be achieved with a smaller number of EA (10 or more) than for the topsoil. With a total sample size of 100 the target is achieved with 50 EA per district, and with a total sample size of 400 the target is achieved with 10 or more EA per district. At the sampling intensities of the UNPS and UHIS surveys, the standard error for the BLUP of subsoil pH varies from 0.06 to 0.09, giving wider confidence intervals than the target.

In the case of soil pH in Uganda we can see some potential for trade-off between the total sample size, and the number of EA which are sampled. For example, for subsoil pH (Fig. 4b) the standard error of the E-BLUP is about the same for a total sample size of 100 over 25 EA per District and for a total sample size of 200 with 10 EA per district. The best solution will depend on the marginal cost of adding additional analyses and how it compares with the cost of adding a new EA to the survey.

Because data on the SOC were analysed on the log-scale, the plots in Fig. 4(c,d) show the width of the back-transformed interval of ±1 standard error, assuming estimates at the global mean, a target width of 0.2 SOC % g/g was specified. For any given sample scheme the uncertainty for topsoil SOC is somewhat larger than for subsoil SOC. The primary difference from the results for soil pH is that there is less scope for trade off between the number of EA and the number of HH

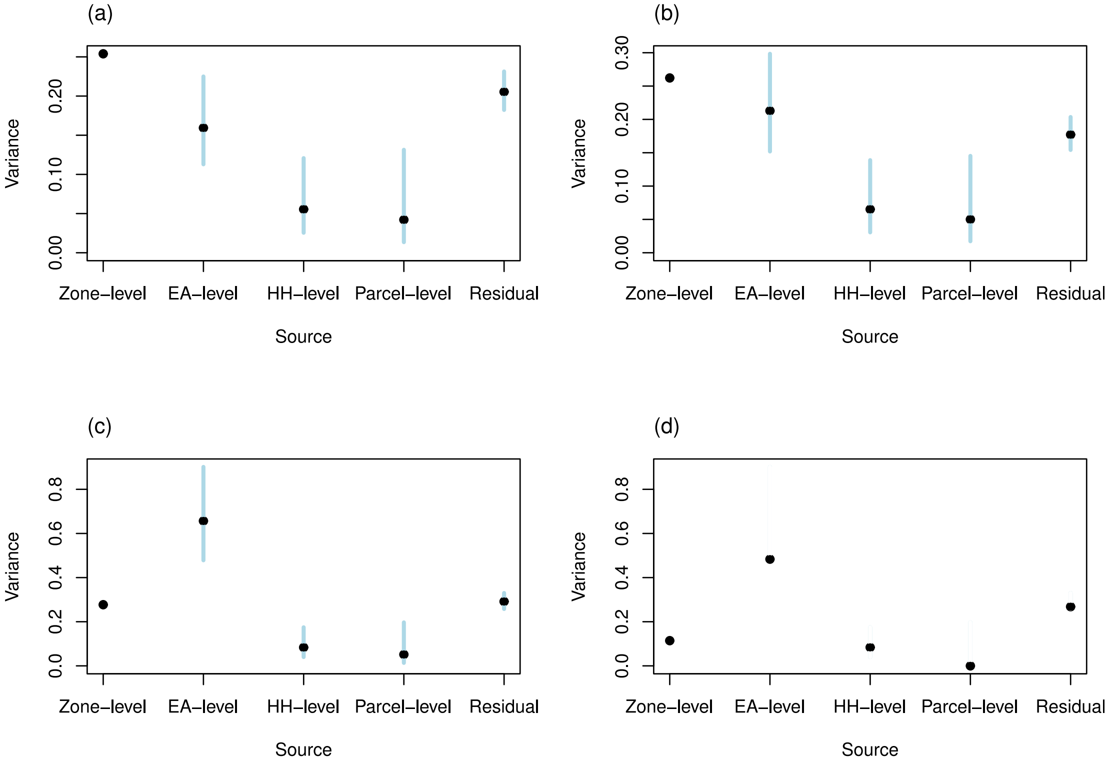

Fig. 3. Ethiopia data: variance components for (a) topsoil pH (b) subsoil pH (c) log topsoil SOC and (d) log subsoil SOC with 95% confidence intervals (not found for subsoil SOC).

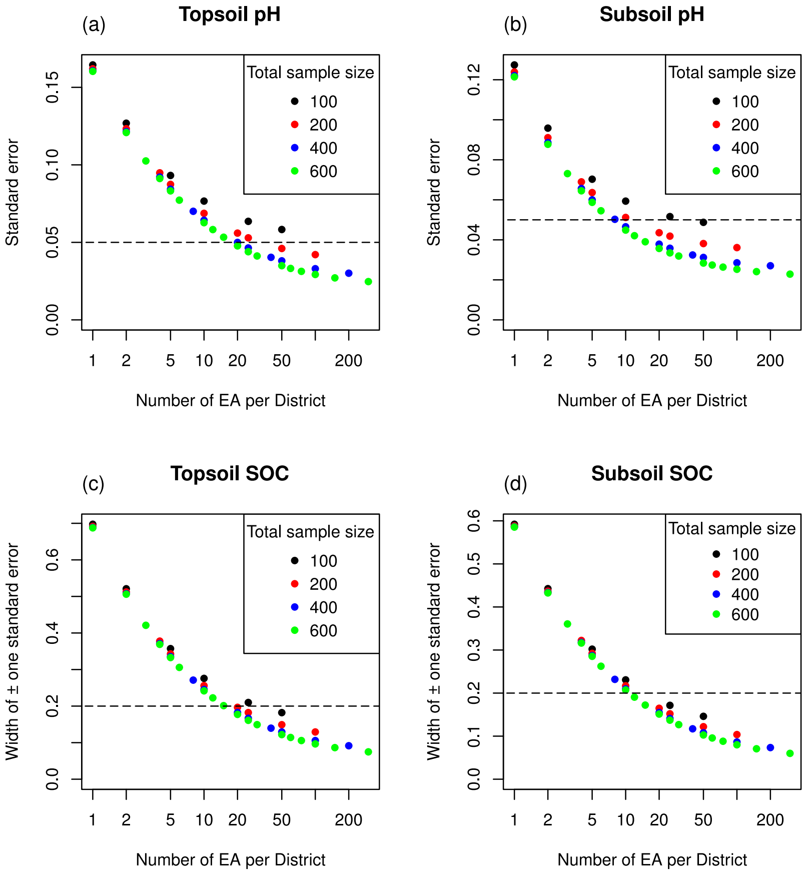

Fig. 4. Uganda: (a) (b) standard error of the E-BLUP of a district mean for (a) topsoil pH (b) subsoil pH plotted against the number of EA per district for different designs. (c) (d) Width of ±1 standard error (original scale) for district mean for (c) topsoil SOC (d) subsoil SOC plotted against the number of EA per district for different designs.

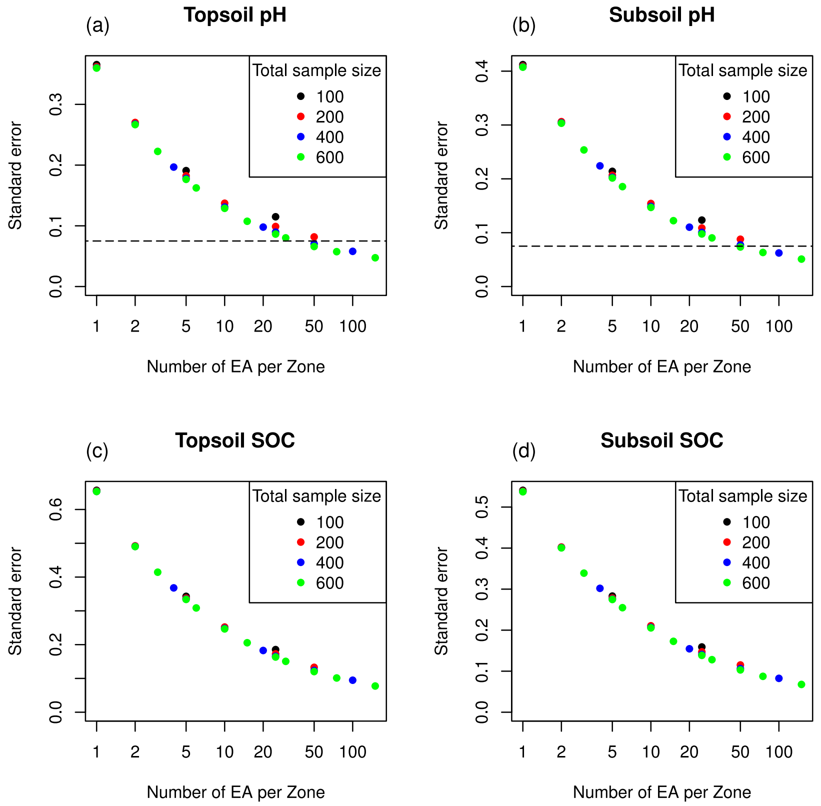

Fig. 5. Ethiopia: Standard error of the E-BLUP of a zone mean for (a) topsoil pH (b) subsoil pH (c) topsoil SOC (d) subsoil SOC plotted against the number of EA per zone for different designs.

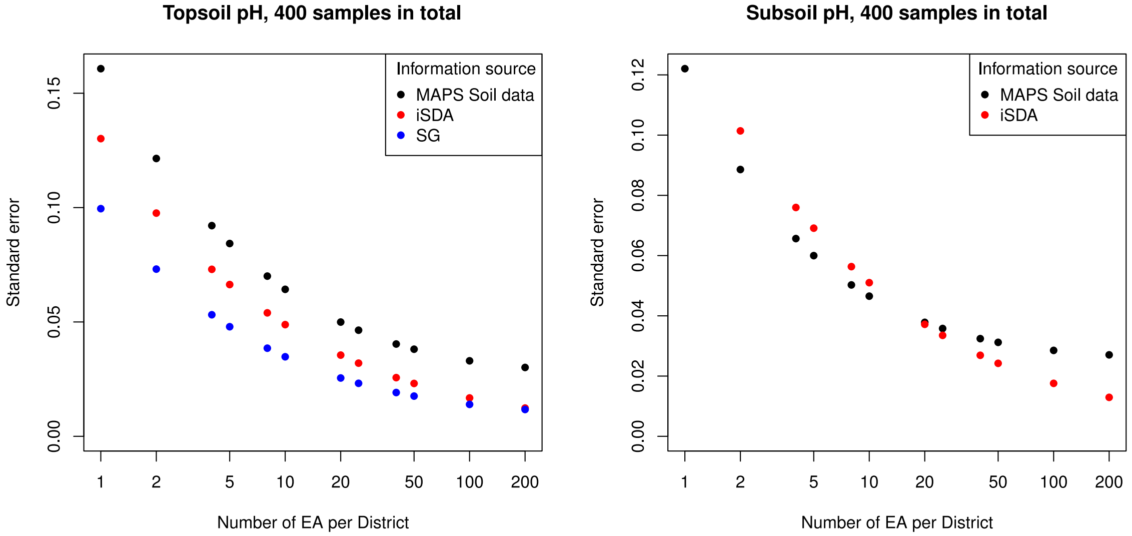

Fig. 6. Uganda: Standard error for district mean soil pH (topsoil left, subsoil right) with different designs with a total sample of 400, based on variance components from different sources: MAPS soil data, iSDA values or SG (Soil Grids) values.

per EA with SOC. This can be attributed to the smaller ratio of the between-HH (residual) variance to the between-EA variance for SOC than for pH. The number of EA is the dominant control of the width of the prediction interval of the E-BLUP, and with more than 10 EA the target is achieved for all sample sizes.

At the sampling intensities of the UNPS and UHIS surveys, the width of the interval of ±1 standard errors on the log scale is 0.25 to 0.45 for subsoil SOC, and 0.3 to 0.5 for topsoil SOC, exceeding the target of 0.2.

3.3.2. *Ethiopia, Zone BLUPs*

The standard errors of the E-BLUPS for Zone means in Ethiopia (Fig. 5) show little dependence on the total sample size, given the number of sampled EAs. This is true for both variables and depths, and can be attributed to the relatively small variance components at HH and parcel level. Soil pH is more variable in Ethiopia than in Uganda, and a standard error of 0.05 pH units is not achieved with the sample designs illustrated in Fig. 5. A wider target standard error of 0.075 pH units, giving a 95% confidence interval of about ±0.15 pH units is achieved with about 50 EA sampled per Zone at both depths.

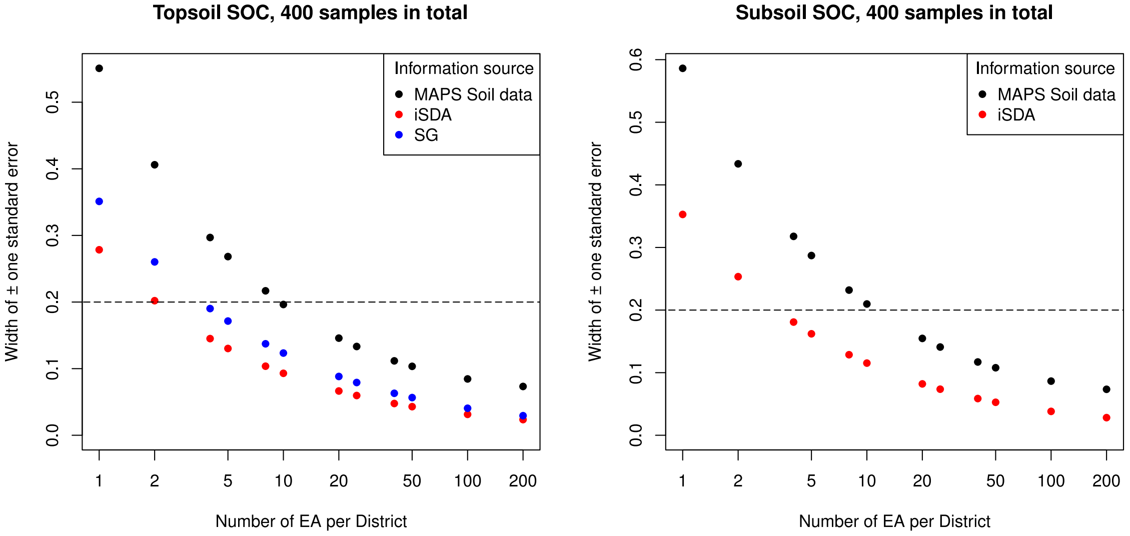

Fig. 7. Uganda: Width of ±1 standard error for district mean SOC (topsoil left, subsoil right) with different designs with a total sample of 400, based on variance components from different sources: MAPS soil data, iSDA values or SG (Soil Grids) values.

Sample intensity for HH surveys in Ethiopia range from 2–3 EA per Zone (ESS) and 6–24 EA per Zone (Ag-SS). At these intensities the standard error for the BLUP of topsoil pH varies from 0.1 to 0.28, 0.1–0.3 (subsoil) giving prediction intervals up to ±0.6 pH units.

A specific target standard error for SOC is not considered, but note that a confidence interval of about ±0.5 SOC % g/g would be achieved at both depths with a sample design with 5 EA per Zone. At EA sampling intensities of ESS and Ag-SS, the standard error of the Zone BLUP for SOC ranges from 0.2 to 0.5 (topsoil) and 0.15 to 0.4 (subsoil).

3.4. Use of DSM products to estimate variance components at different levels of the HH survey

There is little evidence that the variation among the observations in the MAPS 2015 and LASER surveys with respect to soil pH and SOC is well-represented by the corresponding iSDA or Soil Grids data (Figures A13–A16), the closest relationship being for the Soil Grids SOC in Ethiopia (A16). A general discussion of the validity of these DSM products is outside the scope of this paper, but these results do indicate that, at least at present, this information cannot be treated as a substitute for direct measurement of soil properties. The predictions show a marked shrinkage, with much less variation than is seen in the corresponding measurements. For Uganda topsoil pH (Fig. 6a) and SOC at both depths (Fig. 7), the standard errors of the BLUPs based on variance components for DSM predictions are notably smaller than those based on the actual soil measurements. Smaller differences are seen for subsoil pH (Fig. 6b). A similar picture is seen in Ethiopia (Figures A17, A18), particularly for SOC. These results are consistent with the shrinkage effect seen in the scatter plots (Figures A13–A16), and our expectation that DSM predictions would not reproduce the variability of soil properties at the scales of interest.

3.5. Use of variograms from point data to estimate variance components at different levels of the HH survey

The parameters for the variograms of soil properties from the GeoNutrition project data sets are shown in Table 2. The final column of the Table shows the median standardized squared prediction error for the variogram model from its cross-validation. For a correct model the expected value of this statistics is 0.45, and the reported values are all within the 95% confidence interval of this.

Table 2
Variogram parameters estimated from GeoNutrition project data. The term θ is the median standardized squared prediction error from the cross-validation of the parameters. The expected value of this statistic with a valid variogram and normal kriging errors is 0.455, the 95% interval for the statistic, given the sample size, is $[0.262, 0.647]$ for the Ethiopia data set and $[0.388, 0.522]$ for the Malawi data set.

Data set	Variable	Uncorrelated variance	Spatially correlated variance	κ	ϕ /km	Median θ
Ethiopia	pH	0.09	0.50	0.5	31.9	0.430
Ethiopia	SOC	0.43	1.22	0.5	27.6	0.412
Malawi	pH	0.20	0.25	0.5	36.8	0.451
Malawi	log SOC	0.20	0.08	0.5	43.3	0.416

Fig. 8 shows the standard errors for the Zone E-BLUP from samples of 400 observations according to different designs based on the LASER survey variance components and on variance components computed from the GeoNutrition variogram. The values are similar, in particular for topsoil SOC. In both cases the standard errors are somewhat larger for the variance components computed from the variogram than directly from the LASER survey data.

Fig. 9 shows the standard errors for the E-BLUP of the District mean for soil pH in Malawi, and the width of the back-transformed ±1 standard error interval for SOC, based on different survey designs using variance components based on the Malawi GeoNutrition variograms. For both properties there is potential trade-off between the number of EA which are sampled and the total sample size. To achieve a prediction of District mean topsoil pH with a confidence interval of about ±1 pH unit a total sample size of 400 with 50 EA sampled per District would suffice. To have a prediction interval for mean SOC no wider than 0.1 SOC % g/g can be achieved with a sample of 600 points and 10 EA per District, or with a sample of 200 points and 50 EA per district.

4. Discussion

These results illustrate how variance components, estimated from a pilot survey, can be used to compute quality metrics for predictions or estimates based on different sample designs. This provides a basis for selecting a design which provides specified information of adequate precision while avoiding oversampling, or for assessing the potential of an existing sampling design, such as a household survey, as a framework for sampling a new set of variables.

It must be emphasized that in this paper we consider the mean value of a variable for a district (Uganda, Malawi) or zone (Ethiopia) as the objective for sampling, and its standard error as the quality measure

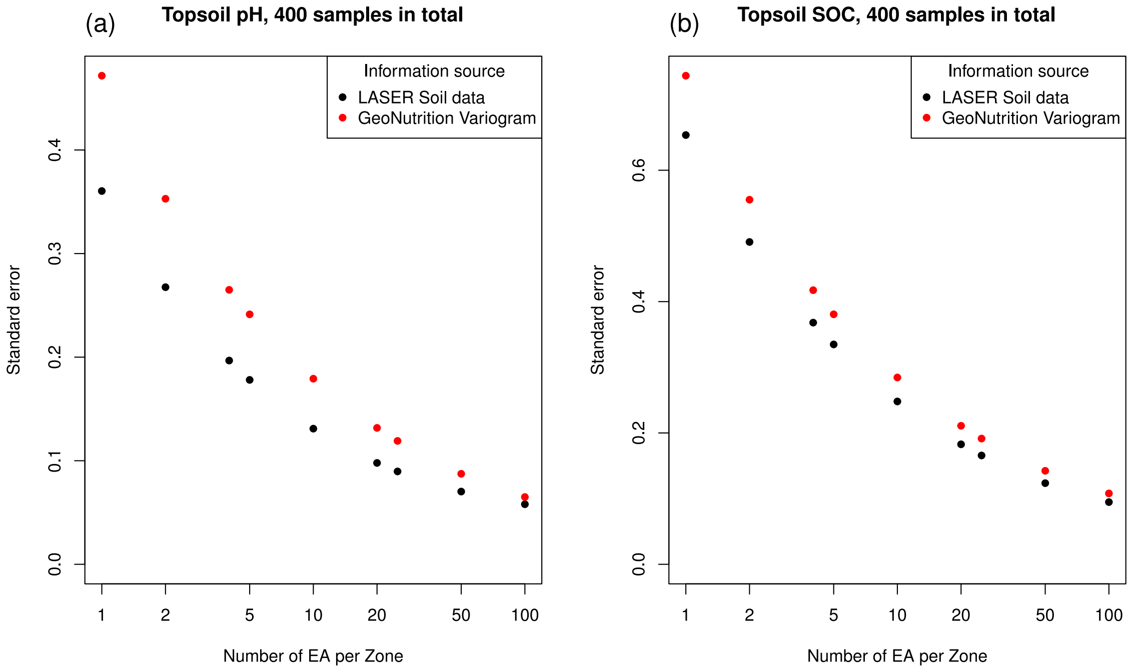

Fig. 8. Ethiopia: standard error for (a) zone mean topsoil pH and (b) zone mean topsoil SOC with different designs with a total sample of 400, based on variance components from LASER soil data and the GeoNutrition variogram.

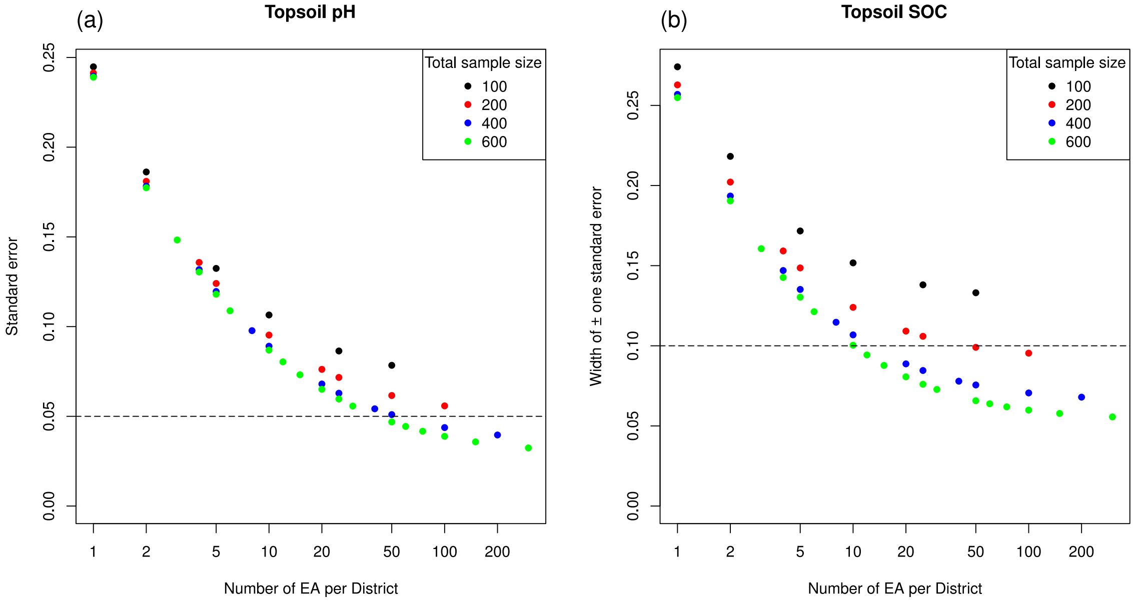

Fig. 9. Malawi: (a) standard error for district mean topsoil pH and (b) Width of ±1 standard error of district mean topsoil SOC on original scale with different designs, based on variance components the GeoNutrition variogram.

for assessing a proposed sampling design. These were selected for illustrative purposes. The results presented here would support a decision about sampling for this objective, but that is not the only objective and associated quality measure which might be considered when supplementing a household survey with soil measurements. For example, one might be interested in mean values for groups of enumeration areas, or a mean value at national or regional scale.

The scale-dependence of soil variation, that is to say the distribution of the variance between different spatial components, affects how sensitive the standard error of an estimate is to different ways of deploying the same sample effort between levels of the sample design. For example, for soil pH in Uganda (Fig. 4a,b), the standard error of an estimate depends both on the total number of EA sampled per district but also the numbers of samples within each EA. This was also found for both variables with variance components extracted from the variograms for the Malawi data (Fig. 9). By contrast, for variables such as SOC in Ethiopia, there is very little difference between the SE for different sample sizes at some fixed number of EA. In the Ethiopian setting the number of EA is the dominant factor determining the SE of a prediction so smaller total sample sizes may be acceptable provided sufficient EA are sampled. The marginal cost of an additional EA in a sample design is likely to exceed the marginal cost of an additional HH within a sampled EA substantially, so where there is a potential trade-off between total sample size and the number of EA to be sampled, larger overall samples are likely to be preferred.

The different locations differ with respect to soil variation (see, for example, the log-normal distribution for SOC in Uganda and Malawi, but not in Ethiopia). Furthermore, the magnitude of the variation of a soil property may differ between locations, as may the scale-dependence. The relative magnitude of the different variance components may also differ between properties in the same region. Note, for example, how the District variance component is a substantial contribution to the variance of SOC in the MAPS 2015 survey from Uganda, but is small for soil pH there. The relatively small contribution of HH and Parcel-level variance components in the LASER survey data explains why total number of EA is the dominant factor influencing the E-BLUP standard errors for Ethiopia.

This has implications for sample requirements. For example, to estimate Zone/District topsoil pH with a standard error no larger than 0.05 requires 400 samples over 20 EA per district in Uganda (Fig. 4a), more than 600 samples in Ethiopia (Fig. 5a) and 400 samples over 50 EA in Malawi. To obtain an estimate of the mean topsoil SOC with the width of ± 1 standard error no greater than 0.2 requires 200 or more samples in 10 EA in Uganda, but 200 samples over 2 EA per district in Malawi. This highlights the importance of obtaining local information for sample planning.

The question remains whether HH surveys are an appropriate framework for soil sampling. In most cases the number of EA used in surveys would constrain the standard errors of E-BLUPs, or the width of prediction intervals to larger values than the targets selected here. However, the latter are essentially arbitrary. The question of how to specify the required precision of an estimate or prediction in statistical terms, specifically how stakeholders with different interests (economists, nutritionists, agriculturalists) might decide what constitutes a sufficiently narrow prediction interval, remains an open one (Lark et al., 2022). Although the confidence interval or prediction interval is a standard measure of uncertainty attached to an estimation or prediction, Chagumaira et al. (2021) found that varied end users did not generally find them easy to interpret alongside outputs, and similarly did not interpret them effectively as quality measures for the design of spatial surveys (Chagumaira et al., 2023). It is possible, for example, that a standard error for District mean pH (topsoil) in the range 0.08 to 0.13, as indicated for current HH surveys in Uganda, could be sufficient for decisions on potential benefits of liming. We therefore cannot conclude, without further stakeholder engagement, whether the indicated precisions of soil information based on HH surveys would meet user requirements.

The differences in scale-dependence, and absolute magnitude of variance components between locations and soil properties, show that generalized rules of thumb for incorporating soil sampling into household surveys are unlikely to result in estimates or predictions of similar precision between regions. If pilot soil data, collected in household surveys such as MAPS 2015 or LASER, are not available then one might base estimates of the necessary variance components on variograms estimated from point data collected in the region of interest in spatial surveys such as those reported by Gashu et al. (2021) or from legacy point data such as the WoSIS data sets (Batjes et al., 2017).

In this study we noted that the sample weights (in so far as these could be approximated post hoc) did not appear to affect the variance components of relevance to the sampling scenarios considered here. However, this might not hold in other circumstances, as the effects of a complex sampling design are known to depend on the variables which are sampled and the estimators which are used Pfeffermann (1993). Therefore, we would recommend that comparable pilot data are analysed with and without sample weights in a similar way.

There might be scope to improve the precision of estimates of soil properties from samples collected in household surveys by model-assisted estimation with appropriate covariates (Särndal et al., 2003). This would be a matter for further research, but we note that the DSM gridded information we examined here seems unlikely to be useful for this purpose. Other covariates such as terrain variables or remote sensor data might be more suitable.

Finally, we note that guidelines on sound design for estimating baseline values of soil properties are not necessarily appropriate for monitoring change in soil over time. This is because the spatial scales at which the change processes dominate might not be comparable with those scales which are most important in the baseline variation (Lark, 2009). If sampling is primarily for monitoring purposes then, ideally, this would be based on resampling of an initial pilot survey to characterize the variability of soil change.

5. Conclusions

The decision to include an additional variable in a household survey, or to integrate direct measurements outwith the survey's original scope, is not a neutral choice, and has consequences. The most straightforward of these is the additional cost imposed on the sponsor of the survey, through the additional labour and time costs of collection of a new variable which, in the case of soil data, includes the laboratory costs entailed in making the measurements. This has been recognized in medical surveys, adding physical measures and the collection of specimens to a survey, as well as adding logistical costs (and additional complexities to the consent procedure), also increases the burden on the survey subject and this may increase the rate of refusal among participants (Boyle et al., 2021), which is a potential source of bias.

For this reason we should carefully evaluate the potential value of information obtained by adding variables to household surveys, and ensure that the number of households for which the additional measurements are made is sufficient to provide information of adequate quality but not excessive, representing unjustified additional sampling costs. The results presented in this study regarding the quality of particular sample estimates are restricted to district or zone means. For this specific objective, the sampling requirement may be substantial; consider, for example, the number of EA per zone specified to achieve a confidence interval of ± 0.15 pH units for the zone mean topsoil or subsoil, this is around 50 EA per zone, which is more than would normally be sampled. However, this is not true for all variables in all the settings we consider. Furthermore, sampling might be feasible for some other objective, for example to estimate the mean value of soil pH for some specified set of EA.

We have shown how past survey data, either collected according to the household survey design, or according to a spatial design, can be used to estimate the variance components of variables of interest, and from this the estimation variance of particular sample means of interest. This will allow informed decisions to be made on the supplementation of household survey protocols with additional observations of the soil, rather than simply incurring this cost, imposing it on participants, and only discovering limitations on the value of resulting elements *post hoc*.

CRediT authorship contribution statement

R.M. Lark: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **L. Mlambo:** Writing – review & editing, Software, Formal analysis. **H. Pswarayi:** Writing – original draft, Formal analysis. **D. Zardetto:** Writing – review & editing, Data curation, Conceptualization. **S. Gourlay:** Writing – original draft, Project administration, Data curation, Conceptualization.

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 article was produced with financial support from the 50 × 2030 Initiative to Close the Agricultural Data Gap, a multi-partner program that seeks to bridge the global agricultural data gap by transforming data systems in 50 countries in Africa, Asia, the Middle East and Latin America by 2030. For more information on the Initiative, visit https://www.50x2030.org/

The authors thank the Ethiopia Statistical Service and the Uganda Bureau of Statistics for their collaboration with the World Bank on the implementation of the methodological studies underlying this research, as well as the World Agroforestry Centre for their role in the soil collection and analysis.

Contributions by HP and LM were funded through by the Bill & Melinda Gates Foundation, United States [INV-002855]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geoderma.2024.117148.

Data availability

The survey data from Ethiopia and Uganda were used by permission of the Ethiopia Statistical Service and the Uganda Bureau of Statistics, and so the authors of this paper are not in a position to share them. The variance components estimated from the data, and the R codes used to compute the standard errors of the E-BLUP for Zone or District means are available via the following link https://github.com/rmlark/Soil-Sampling-in-Household-Surveys

Data from the GeoNutrition project are available at https://doi.org/10.6084/m9.figshare.15911973.

References

Ateku, D., 2021a. Sample Analysis using Bruker Alpha Fourier Transform Mid-Infrared Spectrometer. Standard Operating Procedure SOP 002, World Agroforestry Centre, Nairobi, Kenya, https://www.worldagroforestry.org/sites/agroforestry/files/SOP%20for%20sample%20analysis%20on%20Bruker%20Alpha%20Spectrometer_0.pdf.

Ateku, D., 2021b. Sample Analysis using Bruker Multipurpose Analyzer (MPA) Fourier Transform Near-Infrared Spectrometer. Standard Operating Procedure SOP 006, World Agroforestry Centre, Nairobi, Kenya, https://www.worldagroforestry.org/sites/agroforestry/files/SOP%20for%20sample%20analysis%20on%20Bruker%20Multi-Purpose%20Analyzer%20MPA.pdf.

Ateku, D., Chacha, R., 2021. Samples Reception, Processing, Log-in, Shipping, Archiving and Disposal. Standard Operating Procedure SOP 001, World Agroforestry Centre, Nairobi, Kenya, https://www.worldagroforestry.org/sites/agroforestry/files/SOP%20for%20Sample%20Reception%2C%20Processing%2C%20Log-in%2C%20Shipping%2C%20Archiving%20and%20Disposal.pdf.

Bailey, P., Webb, B., Kelley, C., Nguyen, T., Huo, H., 2023. WeMix: Weighted mixed-effects models using multilevel pseudo maximum likelihood estimation. R package version 4.0.0, https://CRAN.R-project.org/package=WeMix.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., de Mendes, J., 2017. WoSIS: Providing standardised soil profile data for the world. Earth Syst. Sci. Data 9, 1–14. http://dx.doi.org/10.5194/essd-9-1-2017.

Boyle, J., Berman, L., Dayton, J., Iachan, R., Jans, M., ZuWallack, R., 2021. Physical measures and biomarker collection in health surveys: propensity to participate. Res. Soc. Admin. Pharm. 17, 921–929.

Brys, G., Hubert, M., Struyf, A., 2003. A comparison of some new measures of skewness. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in Robust Statistics. Physica-Verlag, Heidelberg, pp. 98–113.

Carle, A.C., 2009. Fitting multilevel models in complex survey data with design weights: recommendations. BMC Med. Res. Methodol. 9 (49).

Carletto, C., Gourlay, S., 2019. A thing of the past? Household surveys in a rapidly evolving (agricultural) data landscape: Insights from the LSMS-ISA. Agricult. Econ. 50 (S1), 51–62.

Central Statistical Agency, 2017. Ethiopia land and soil experimental research (LASER) study basic information document. Available at: https://microdata.worldbank.org/index.php/catalog/2671/download/40949.

Chagumaira, C., Chimungu, J.C., Gashu, D., Nalivata, P.C., Broadley, M.R., Milne, A.E., Lark, R.M., 2021. Communicating uncertainties in spatial predictions of grain micronutrient concentration. Geosci. Commun. 4, 245–265.

Chagumaira, C., Chimungu, J.C., Nalivata, P.C., Broadley, M.R., Milne, A.E., Lark, R.M., 2023. Planning a geostatistical survey to map soil and crop properties: eliciting sampling densities. Geosci. Commun. Discuss. http://dx.doi.org/10.5194/gc-2023-1, [preprint]. in review.

Diggle, P.J., Ribeiro, P.J., 2007. Model-based Geostatistics. Springer-Verlag, New York.

Gashu, D., Nalivata, P.C., Amede, T., Ander, E.L., Bailey, E.H., Botoman, L., Chagumaira, C., Gameda, S., Haefele, S.M., Hailu, K., Joy, E.J.M., Kalimbira, A.A., Kumssa, D.B., Lark, R.M., Ligowe, I.S., McGrath, S.P., Milne, A.E., Mossa, A.W., Munthali, M., Towett, E.K., Walsh, M.G., Wilson, L., Young, S.D., Broadley, M.R., 2021. Cereal micronutrient quality varies geospatially in Ethiopia and Malawi. Nature 594, 71–76.

Giller, K.E., Delaune, T., Silva, J.V., et al., 2021. Small farms and development in sub-saharan africa: Farming for food, for income or for lack of better options? Food Secur. 13, 1431–1454.

Gödecke, T., Stein, A.J., Qaim, M., 2018. The global burden of chronic and hidden hunger: trends and determinants. Glob. Food Secur. 17, 21–29.

Gourlay, S., Kilic, T., Lobell, D., 2019. A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale-productivity relationship in uganda. J. Dev. Econ. 141, 102367. http://dx.doi.org/10.1016/j.jdeveco.2019.102376.

Grafström, A., Schelin, L., 2014. How to select representative samples. Scand. J. Stat. 41, 277–290.

Henderson, C.R., Kempthorne, O., Searle, S.R., von Krosigk, C.M., 1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics 15, 192–218.

Hengl, T., Miller, M.A.E., Križan, J., et al., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11 (6130), http://dx.doi.org/10.1038/s41598-021-85639-y.

iSDA, 2024. Open access soil data for Africa. http://dx.doi.org/10.17605/OSF.IO/A69R5.

Journel, A.G., Huijbregts, Ch.J., 1978. Mining Geostatistics. Academic Press, London & New York.

Kumssa, D.B., Mossa, A.W., Amede, T., et al., 2022. Cereal grain mineral micronutrient and soil chemistry data from geonutrition surveys in ethiopia and malawi. Scientific Data 9, 1–12. http://dx.doi.org/10.1038/s41597-022-01500-5.

Lal, R., Bouma, J., Brevik, E., Dawson, L., Field, D.J., Glaser, B., Hatano, R., Hartemink, A.E., Kosakig, T., Lascelles, B., et al., 2021. Soils and sustainable development goals of the united nations: An international union of soil sciences perspective. Geoderma Reg. 25, e00398.

Lark, R.M., 2003. Two robust estimators of the cross-variogram for multivariate geostatistical analysis of soil. Eur. J. Soil Sci. 54, 187–201.

Lark, R.M., 2009. Estimating the regional mean status and change of soil properties: two distinct objectives for soil survey. Eur. J. Soil Sci. 60, 748–756.

Lark, R.M., Chagumaira, C., Milne, A.E., 2022. Decisions, uncertainty and spatial information. Spat. Statist. 50, 100619.

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.

Matheron, G., 1962. Traité de Géostatistique Appliqué, Tome 1. Memoirs du Bureau de Recherches Géologiques et Minières, Paris.

Pfeffermann, D., 1993. The role of sampling weights when modeling survey data. Internat. Statist. Rev. 61, 317–337.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rawlins, B.G., Lark, R.M., O'Donnell, K.E., Tye, A., Lister, T.R., 2005. The assessment of point and diffuse soil pollution from an urban geochemical survey of Sheffield, England. Soil Use Manage. 21, 353–362.

Särndal, C.-E., Swensson, B., Wretman, J., 2003. Model Assisted Survey Sampling. Springer-Verlag, New York.

Singh, S.K., Sharma, S.K., Rana, M.J., Porwal, A., Dwivedi, L.K., 2022. Effectiveness of modular approach in ensuring data quality in large-scale surveys: Evidence from national family health survey - 4 (2015–2016). SSM Popul. Health 19, 101254.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Mass. and London.

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

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd Edition. John Wiley & Sons, Chichester.