Patron: Her Majesty The Queen

ROTHAMSTED RESEARCH

Rothamsted Research
Harpenden, Herts, AL5 2JQ

Telephone: +44 (0)1582 763133
Web: http://www.rothamsted.ac.uk/

Rothamsted Repository Download

A - Papers appearing in refereed journals

Alemu, R., Gelaw, A. M., Gashu, D., Tafere, K., Mossa, A. W., Bailey, E. H., Masters, W. A., Broadley, M. and Lark, R. M. 2022. Sub-sampling a large physical soil archive for additional analyses to support spatial mapping; a pre-registered experiment in the Southern Nations, Nationalities, and Peoples Region (SNNPR) of Ethiopia. *Geoderma.* 424 (15 Oct), p. 116013. https://doi.org/10.1016/j.geoderma.2022.11601

The publisher's version can be accessed at:

- https://doi.org/10.1016/j.geoderma.2022.11601

The output can be accessed at: https://repository.rothamsted.ac.uk/item/98983/sub-sampling-a-large-physical-soil-archive-for-additional-analyses-to-support-spatial-mapping-a-pre-registered-experiment-in-the-southern-nations-nationalities-and-peoples-region-snnpr-of-ethiopia.

© 28 June 2022, Please contact library@rothamsted.ac.uk for copyright queries.

27/07/2022 14:49 repository.rothamsted.ac.uk library@rothamsted.ac.uk

Rothamsted Research is a Company Limited by Guarantee
Registered Office: as above. Registered in England No. 2393175.
Registered Charity No. 802038. VAT No. 197 4201 51.
Founded in 1843 by John Bennet Lawes.

Geoderma 424 (2022) 116013

Contents lists available at ScienceDirect

Geoderma

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

Sub-sampling a large physical soil archive for additional analyses to support spatial mapping; a pre-registered experiment in the Southern Nations, Nationalities, and Peoples Region (SNNPR) of Ethiopia

R. Alemu [a], A.M. Gelaw [b], D. Gashu [c], K. Tafere [d], A.W. Mossa [e], E.H. Bailey [e], W.A. Masters [a,f], M.R. Broadley [e,g], R.M. Lark [e,*]

[a] Tufts University, The Fletcher School and Department of Economics, 177 College Avenue, Medford, MA 02155, USA
[b] Ethiopian Agricultural Transformation Agency, Addis Ababa, Ethiopia
[c] Centre for Food Science and Nutrition, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia
[d] The World Bank, Development Research Group, 1818 H Street Northwest, Washington D.C., USA
[e] School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK
[f] Tufts University, Friedman School of Nutrition Science and Policy, 150 Harrison Avenue, Boston, MA 02111, USA
[g] Rothamsted Research, West Common Harpenden, Hertfordshire, AL5 2JQ, UK

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Soil archive
Sub-sampling
Selenium
Iodine
Pre-registered experiment

2000 MSC:
0000
1111

ABSTRACT

The value of physical archives of soil material from field sampling activities has been widely recognized. If we want to use archive material for new destructive analyses to support a task, such as spatial mapping, then an efficient sub-sampling strategy is needed, both to manage analytical costs and to conserve the archive material. In this paper we present an approach to this problem when the objective is spatial mapping by ordinary kriging. Our objective was to subsample the physical archive from the Ethiopia Soil Information System (EthioSIS) survey of the Southern Nations, Nationalities and Peoples Region (SNNPR) for spatial mapping of two variables, concentrations of particular fractions of selenium and iodine in the soil, which had not been measured there. We used data from cognate parts of surrounding regions of Ethiopia to estimate variograms of these properties, and then computed prediction error variances for maps in SNNPR based on proposed subsets of the archive of different size, selected to optimize a spatial coverage criterion (with some close sample pairs included). On this basis a subsample was selected.

This is a preregistered experiment in that we have proposed criteria for evaluating the success of our approach, and are publishing that in advance of receiving analytical data on the subsampled material from the laboratories where they are being processed. A subsequent short report will publish the outcome. The use of preregistered trials is widely recommended and used in areas of science including public health, and we believe that it is a sound strategy to promote reproducible research in soil science.

1. Introduction

Soil information is required to support, among other activities, agricultural development, environmental protection and the improved management of soil and crop systems to ensure that food provides sufficient micronutrients (Gashu et al., 2021). Much of the cost of acquiring soil information is associated with field work and sampling (Lark and Knights, 2015). For this reason, it is good practice wherever possible to maintain a physical archive of soil material from any significant field survey. This has been costed into proposed national-scale soil monitoring schemes (e.g. Black et al., 2008). The existence of a physical archive of soil from past sampling has enabled advances such as the development of the RothC soil carbon model, using data from soil samples from the long-term experiments at Rothamsted Experimental Station (Jenkinson and Rayner, 1977). The re-analysis of soil samples from the National Soil Inventory of England and Wales provided data on elemental concentrations in these soils, determined by X-ray fluorescence spectrometry. A total of 53 elements were measured in this way, in contrast with the total of 17 measured from the samples at the time of the original survey (Rawlins et al., 2012).

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

https://doi.org/10.1016/j.geoderma.2022.116013
Received 21 March 2022; Received in revised form 13 June 2022; Accepted 15 June 2022
Available online 28 June 2022
0016-7061/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

When a problem arises which can be tackled by making new measurements on soil in a physical archive, the question follows: which specimens in the archive should be used? If possible we might use the whole archive, but there are two problems with this. First, the soil archive, as a snapshot of the soil over a limited time window, is irreplaceable. Any destructive analysis therefore reduces its future value. Second, the costs of new analyses may be large, so it makes sense to select a subset of material for examination depending on the question in hand. In the case of spatial mapping by ordinary kriging the utility of the spatial predictions of a soil property, as measured by the kriging variance, depends on the spatial dependence of the variable of interest and the spatial distribution of the sample (McBratney and Webster, 1981). This means that an appropriate subsample of the physical archive can be selected for new soil analyses to allow spatial prediction of the new variable by ordinary kriging, if we know the original locations of all the specimens, and have a statistical model of the spatial dependence of the variable. This model can be derived from data in neighbouring or otherwise comparable locations. Similarly, if the soil data are combined with exhaustive covariates in a linear mixed model for spatial prediction by the empirical best linear unbiased predictor (equivalent to kriging with an external drift), then the prediction error variance can be computed in advance for locations where the covariates are measured, given the original locations of points in the physical sample, and an approximation to the parameters of the linear mixed model.

In this study we had the task of selecting a subset of samples from the physical archive of the Ethiopian Soil Information System (EthioSIS) to allow spatial mapping by ordinary kriging of soil variables not measured in the orginal project, but which can be determined by new chemical analyses of the archived material. This study is focussed in the Southern Nations, Nationalities and Peoples Region of Ethiopia (SNNPR). The EthioSIS survey, conducted by the Ethiopian government's Ministry of Agriculture and the Ethiopian Agricultural Transformation Agency (ATA), gathered and analyzed soil samples from each of Ethiopia's 18,000 agricultural local administrations (kebeles), cataloguing over 100,000 samples. The primary objective of the project was to develop maps of soil health and fertility which would facilitate location-specific recommendations on fertilizer blending and to support decisions on land use and the suitability of different crops and varieties.

Recent studies in sub-Saharan Africa, including Ethiopia, have shown that the mineral micronutrient status of human populations, staple crops and the properties of soils including measures of micronutrient concentration show related, spatially dependent, geographical distributions, (Ligowe et al., 2020; Gashu et al., 2021; Belay et al., 2020; Belay et al., 2021). This implies that information on such properties of the soil could help to direct interventions to tackle mineral micronutrient deficiency, particularly in countries where deficiency is prevalent, and many people depend on locally-grown staple crops. For this reason, information on micronutrient status of soils could be useful, alongside more conventional properties of agronomic relevance. In Ethiopia iodine and selenium deficiencies have been reported to be prevalent, which may partly be explained by poor supply from the soil, given the historically small rates of salt-iodization and selenium fortification coupled with heavy dietary dependence on cereals and legumes.

Although soils collected for the EthioSIS project were analyzed to measure the concentrations of several macro- and micro-nutrients, iodine and selenium were not included. Furthermore, there are specific extractions of these elements which are most likely to relate to the crop-available supply of the element and these have not been used in the EthioSIS project (Gashu et al., 2021).

To fill this gap and shed light on the link between soil and human mineral status, a subset of the total sample collected by EthioSIS from Ethiopia's SNNP regional states is to be selected for further examination of the soil iodine and selenium concentrations which are bio-available for plant uptake. This is to be done with particular extractants, described below, and the extractions are to be analyzed using an inductively coupled plasma mass spectrometer (ICP-MS) to determine concentrations of organic iodine and soluble selenium. Our specific objective is to facilitate spatial mapping of these fractions of soil iodine and selenium to identify spatial variations in the supply to crops and local communities.

We propose to select the subsample from the EthioSIS archive for SNNPR for these analyses such that ordinary kriging predictions of the target variables have adequate precision as judged by the expected ordinary kriging variances. This requires a prior variogram model for the soil properties of interest (McBratney and Webster, 1981). We shall use estimated and validated variogram models from existing data in the Amhara, Tigray and Oromia Regions of Ethiopia which were collected as part of the GeoNutrition Project (Gashu et al., 2021). No fieldwork was done in SNNPR for this project. Fig. S1 in the supplementary information shows the regions of interest, and Fig. 1 shows the distribution of sample points from the GeoNutrition Project in Amhara, Tigray and Oromia, and the distribution of EthioSIS sample points in SNNPR.

This is a pre-registered experiment. At the time of submission the data from the analysis of the subsample of the EthioSIS physical archive are not complete or available to the analytical team. This is in line with innovations to improve the reproducibility of science, by proposing in advance the details of an experiment and the criteria on which it will be evaluated. This is widely done in medical science and public health (e.g. Lowe et al., 2020), and increasingly in other domains such as agronomy (Botoman et al., 2020). In the following sections we set out the methodology for selecting a subsample of the EthioSIS physical soil survey in SNNPR, and the criteria by which results from the final analysis will be evaluated to judge how successful the selection of the subset was.

2. Methods

2.1. The EthioSIS soil sampling strategy

Sampling sites were selected in two steps. First, the 97 confluence points in Ethiopia, i.e. the intersection of lines of latitude and longitude at 1-degree intervals, were identified. Second, at each confluence point a 10- × 10-km area was subdivided into 16 tiles, and within each of these tiles 10 sample points were selected at random. Given that the main focus of the EthioSIS project was to develop customized fertilizer recommendation for crop cultivation, the EthioSIS team selected sample sites primarily from areas used for cereal production (80 percent) while the remaining 20 percent of the samples were drawn from areas with capability for crop production in future. Sites that were deemed unsuitable for cereal production were excluded.

A soil sample was collected from the pre-defined primary sampling point and eight additional sub-sampling points located 15 meters away from the center of the primary sampling point. This sampling approach was employed to minimize potential bias that could arise from convenience composite sampling. In line with the effective root depth of most annual crops, samples were collected from a uniform depth of 20 cm using a 1.2-metre auger. Soil samples collected from each of the nine sites were pooled into one plastic bucket. After mixing, a representative sub-sample of 1 kg was retained and placed into a labelled sample bag. The sample label includes the district ID, sample plot ID, date of sampling, sample depth and a unique bar code. Samples were finally shipped to the nearest regional or national soil laboratory for further processing. After analysis the remaining material was preserved in a physical archive. Gelaw et al. (2018) give an overview of the project.

2.1.1. Defining a prediction domain

The objective of this study was to enable predictive mapping of soluble Se and organic I in the cultivated soils of SNNPR after new analyses of a subset of archived soil material from the Region. Note that this objective was consistent with the selection principles for the EthioSIS survey under which sites were selected from land in use for cereal production, or with capability for this use. We defined the domain for the spatial predictions from a 500-m raster grid across SNNPR. At

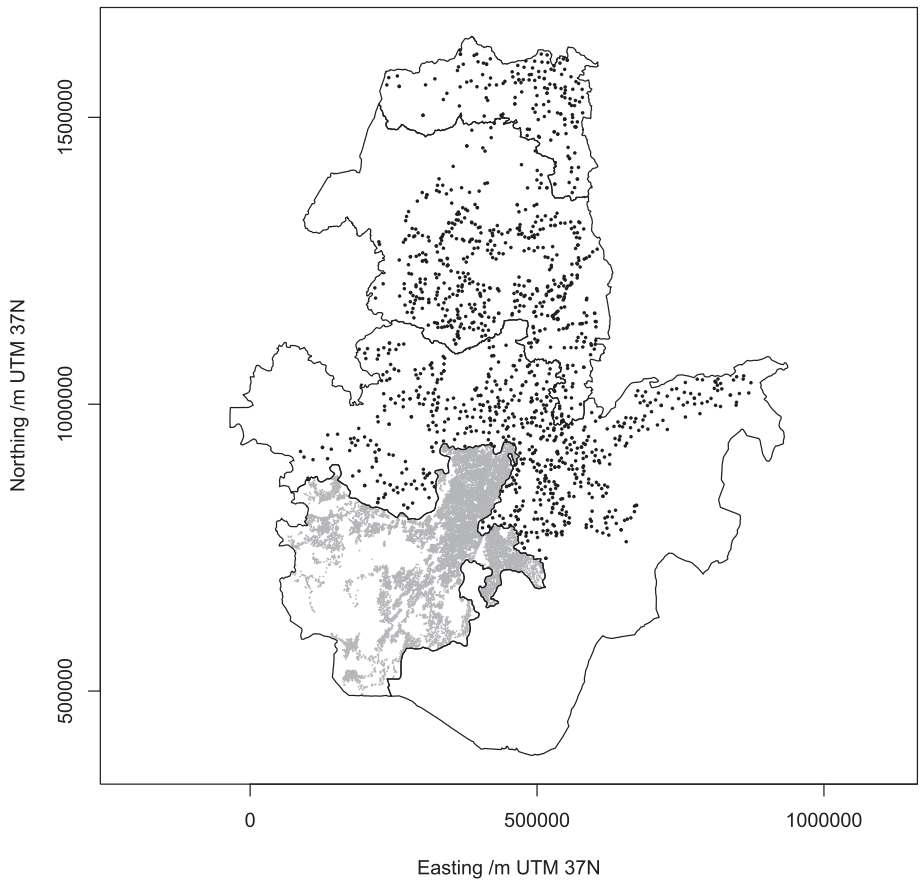

Fig. 1. Map of Tigray, Amhara, Oromia and SNNP Regions of Ethiopia (see Supplementary Fig. 1). The black symbols are GeoNutrition sample sites, the grey symbols are EthioSIS sample sites within SNNP Region.

each of the grid nodes the probability that cropping is the landuse had been computed by the AfSIS project (Walsh et al., 2019). These probabilities were output from a machine learning algorithm with covariates derived from remote sensor data and digital elevation models (AfSIS, 2015) as inputs, and calibration observations made by trained observers using high-resolution satellite imagery. Fig. 2(a) shows these probabilities across SNNPR discretized into three intervals. In Fig. 2(b) the locations of the EthioSIS sample sites in the region are mapped. The Figure shows that the region where the probability of cropping exceeds 0.3 encompasses almost all the EthioSIS sample points, although some of this region is relatively empty of points. On this basis it was decided to treat the raster nodes where probability of cropping exceeds 0.3 as the

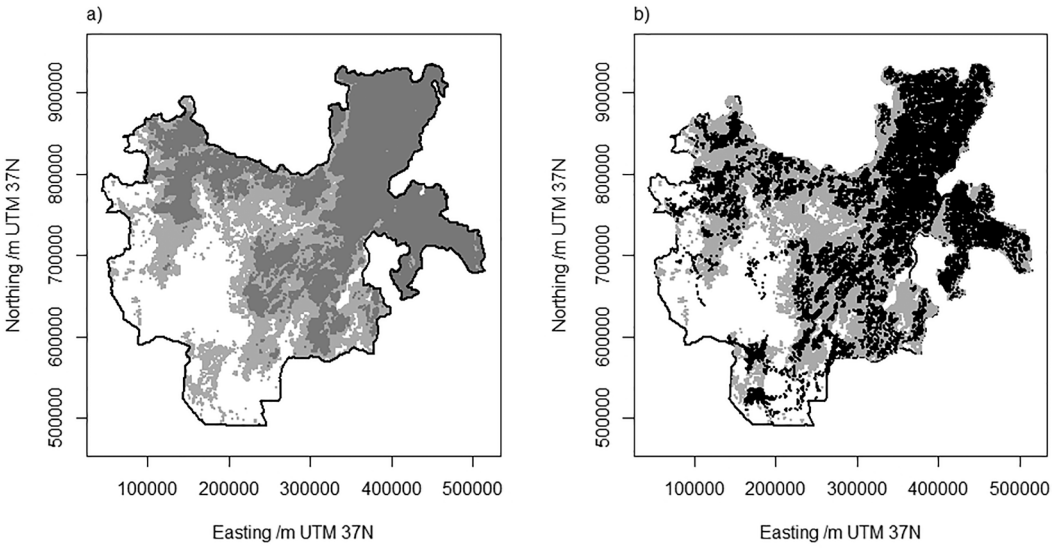

Fig. 2. a). Probability that land is cropped on a grid across SNNPR. White: $P < 0.3$; Light grey: $0.3 \leqslant P < 0.9$; Dark grey: $0.9 \leqslant P$. b). Distribution of sample points across SNNPR (solid black symbols). The grey mask represents where the probability of cropping is greater than 0.3.

domain for purposes of spatial prediction of soil properties. This corresponds to 62% of the area of the region.

2.2. The GeoNutrition data and their analysis

To complete the task of selecting a subset of EthioSIS samples in SNNPR for new analyses to allow ordinary kriging predictions of organic I and soluble Se at points in the prediction domain we require variogram models for the target variables. From these it is possible to compute ordinary kriging variances for the variables based only on the spatial distribution of the selected archive sites. Data on the soil variables were available from the GeoNutrition project from samples collected across Amhara, Tigray and Oromia regions of Ethiopia, but not SNNP, (Gashu et al., 2021). The spatial distribution of these data are shown by the black symbols in Fig. 1.

2.2.1. Collection and chemical analysis

A detailed account of the soil sampling protocol for the GeoNutrition survey is given by Gashu et al. (2021). In summary, the sample locations were selected by balanced spatial sampling with spatial spread using the cube method of Deville and Tillé (2004). Spatial balance means that the mean coordinates of the sample points should be close to the mean coordinates of the sampled domain, and spatial spread means that the points give good spatial coverage avoiding clustering, see Grafström and Schelin (2014). Five soil cores (to depth 150 mm, diameter 50 mm) were collected from within a 100-m^2 circular plot. These were thoroughly mixed and then oven-dried at 40 °C for 24 – 48 h. The soil in each sample was then disaggregated and sieved to pass 2 mm. Subsamples were formed by coning and quartering to produce 150-g aliquots. These were shipped for analysis at the University of Nottingham (for the measurements reported here).

The soil variables we consider here were measured as part of a sequential fractionation procedure adapted from the methods of Mathers et al. (2017) and Shetaya et al. (2012) to extract three fractions of Se and I. Of these the 'soluble' fraction was extracted in 0.01-M KNO_3 and the 'organic' fraction was extracted in 10% tetramethylammonium hydroxide (TMAH). Selenium and I in the extracts were analysed using QQQ-ICP-MS operated in oxygen cell mode with Se mass-shifted from m/z 80 to m/z 96 (SeO^+) to avoid interference from the argon dimer. We refer to the variables as soluble Se and organic I in the remainder of this paper.

2.2.2. Spatial analysis of the GeoNutrition data

As we do not have any data on the target variable from soils of SNNPR, the objective is to obtain variograms from data on those variables in the GeoNutrition data set from adjacent regions of Ethiopia. The GeoNutrition data on the target variables show pronounced North–South trends in both the soluble Se and organic I in the Amhara, Oromia and Tigray data. Figs. S2 and S3 in the supplementary material show the spatial distribution and the quartiles of both variables. In both the larger values are particularly predominant in the southern part of the data set, particularly for organic I. The SNNP Region is south of the sample locations in the GeoNutrition project at longitudes up to about 41 degrees. On inspection of the post plots in Figs. S2 and S3, it was decided to use only the GeoNutrition data from latitudes of 10 degrees or less north, on the grounds that these were more likely to resemble data in SNNPR than samples from further north. Some north–south trend may remain in this restricted subset of the GeoNutrition data. However, this is compatible with ordinary kriging, in which only the local mean is assumed to be constant. Attempting to model a trend would not help with our task, because it could not confidently be extrapolated across SNNPR.

To check this decision further the Soil Reference Groups (World Reference Base, 2007) mapped on the Soil Atlas of Africa (Jones et al., 2013) were extracted for the EthioSIS SNNPR sites, and for the GeoNutrition sample sites with latitudes ⩽10 degrees north. The relative proportions of the different groups in the two subsets are shown in Fig. S4 in the supplementary material. While the compositions of the subsets with respect to the soil groups are not identical they are broadly similar, with Vertisols, Nitosols, Luvisols and Leptosols the commonest groups in both subsets, and Fluvisols, Cambisols, Alisols, Andosols and Phaeozems occurring in smaller proportions. The data on organic I and soluble Se from locations on latitude 10 degrees or less North are referred to as the data subset.

2.2.3. Variogram estimation, modelling and validation

Summary statistics for the data subset on soluble Se and organic I were computed, and are shown in Table 1. Histograms with superimposed box-and-whisker plots, and Q-Q plots of the quantiles of the data against the corresponding normal quantiles were also computed (supplementary Figs. 5–7). The octile skewness in Table 1 is a robust measure of skewness (Brys et al., 2003) based on the asymmetry of the seventh and eighth octile around the median value. A rule of thumb is to consider a transformation for variables with a conventional skewness coefficient outwith the interval $[-1, 1]$ (Webster and Oliver, 2007), but this coefficient can be unduly influenced by outlying values, so the octile skewness is also considered, along with data plots. Rawlins et al. (2005) found that a corresponding rule of thumb for a range of distributions is to consider transformation if the octile skewness is not in the range $[-0.2, 0.2]$. On this basis all further analyses were conducted on the soluble Se data transformed to natural logarithms, and on the untransformed data on organic I.

Empirical variograms for the transformed soluble Se and organic I data subsets were computed for lag-bins using the standard estimator due to Matheron (1962). Distances between observations were computed from their recorded latitude and longitude on a spherical approximation using the distVincentySphere function from the geosphere package of functions for the R platform (Hijmans, 2019; R Core Team, 2020). There was no evidence of directional-dependence in these estimates (Figs. S7 and S8 in the supplementary material), so isotropic variograms were computed. In addition, isotropic variograms were computed with the robust estimators due to Cressie and Hawkins (1980) and Dowd (1984). Although exploratory analysis of the data did not indicate the presence of any marginal outliers, with spatial data one must also consider the possible occurrence of spatial outliers, values which are very unusual in their spatial context, and which might have an undue influence on estimates of the variogram based on the mean-squared difference over different lag distances (Lark, 2000). The robust estimators damp such effects, although in different ways.

All estimated variograms were then fitted with exponential models by weighted least squares, weighting by the number of pair-comparisons within each lag bin (Webster and Oliver, 2007). The exponential model was used because it corresponds to a limiting form of the Matérn variogram model which is positive-definite on the sphere (Gneiting, 2013). The empirical variograms and fitted models are shown in Fig. 3 (Se) and Fig. 4 (I). The fitted models were then assessed by leave-one-out cross-validation. The standardized squared prediction errors (squared prediction errors divided by the kriging variance) were computed from the

Table 1
Summary statistics for soluble Se (Se_{sol}) and organic I (I_{org}) concentration in GeoNutrition soil samples at or south of latitude 10° north.

Variable	Se_{sol}		I_{org}	
Units	μg kg^{-1}	log (μg kg^{-1})	mg kg^{-1}	log (mg kg^{-1})
Mean	3.90	1.17	17.33	2.68
Median	3.14	1.14	17.51	2.86
Q1	2.11	0.75	9.69	2.27
Q3	5.09	1.63	23.26	3.15
SD	2.61	0.61	9.01	0.63
Skewness	2.08	0.03	0.37	0.03
Octile skewness	0.37	0.03	−0.04	−0.36

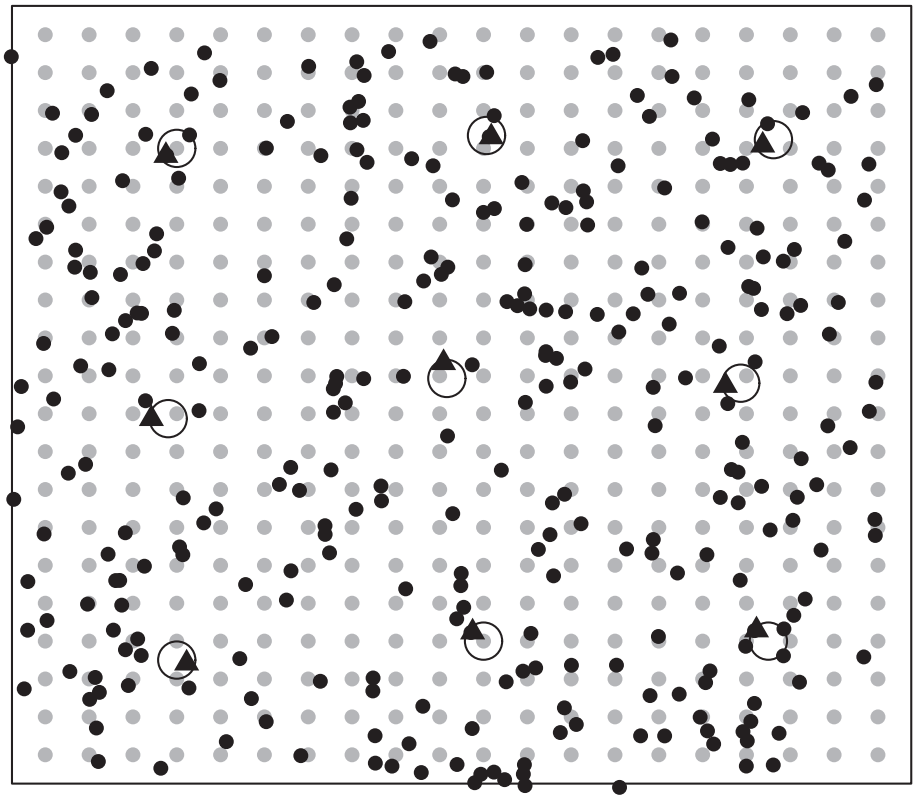

Fig. 5. Illustration of the method for selection of the basic subset of EthioSIS archive samples. The grey discs represent nodes within a subset of the sample domain represented in the Figure. The large circles are the centroids of nine clusters of domain nodes formed by the spcosa package. The black symbols represent the location of samples in the EthioSIS archive, the larger black triangles being the closest point to each of the respective centroids, selected for the subsample.

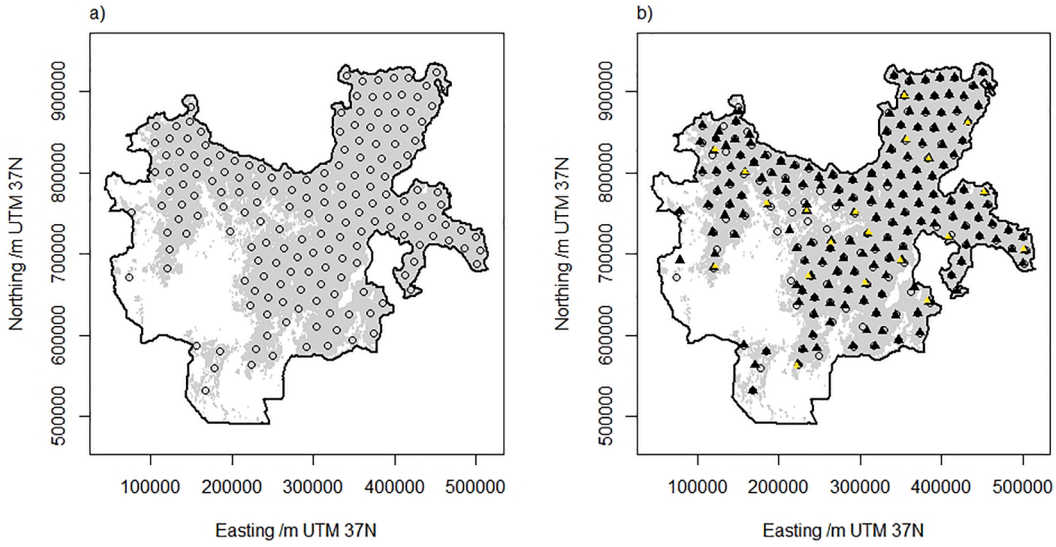

Fig. 6. a) Map of SNNPR showing the sampling domain (grey) where the probability of cropping as the dominant land use equals or exceeds 0.3. The open symbolds show the centroids of 180 polygons formed by *k*-means clustering of the domain points. b) as in Fig. 4a, but with the location of EthioSIS sample sites closest to each sample domain centroid indicated by a black triangle. The yellow triangle near 20 of these locations indicates an EthioSIS sample to be included as a close-paired observation.

cross-validation results. The expected value of the median standardized squared prediction error is 0.455 for normal kriging errors, and we followed Lark (2000) in using this statistic as a diagnostic, and using Matheron (1962)'s estimator if its standardized squared prediction error fell within the 95% confidence interval given the sample size. Otherwise the variogram models fitted to the robust estimates would be considered.

The cross-validation errors for the variogram models both variables, fitted to the estimates obtained by Matheron's estimator, appeared close to normal in distribution (Figs. S9 and S10 in the Supplementary Material), and, as seen in Table 1, the standardized squared prediction errors had median values close to the expectation for a valid model and well-

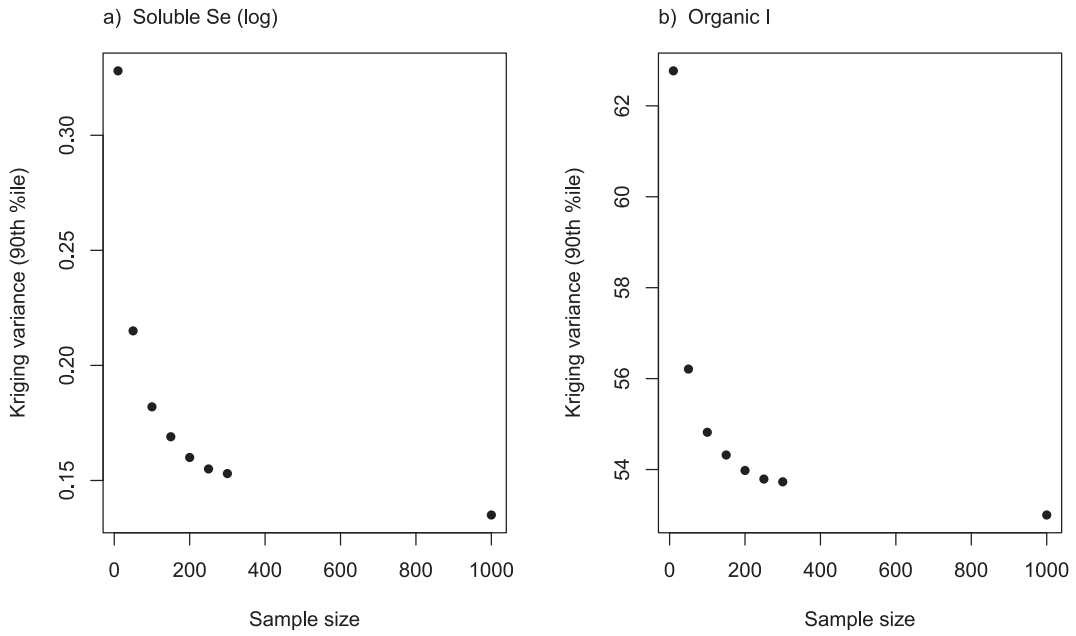

Fig. 7. a) The 90th percentile of kriging variances at a spatially balanced sample of locations in the target domain across SNNPR as a function of the total sample size deployed for a) log of soluble Se and b) organic I.

within the 95% confidence interval. On this basis these models, with parameters in Table 1, were used for all further analyses.

2.3. Sampling objectives and design

For purposes of the analyses all coordinates on EthioSIS sample locations and prediction domain nodes in SNNPR were converted to units of metres relative to the datum of the UTM Zone 57 N projection. This is because the methods described in the following section require locations on rectilinear coordinates. We denote by $\mathscr{S}$ the set of m vectors $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$, where $\mathbf{s}_i \in \mathscr{S}$ denotes the coordinates (UTM37N) of the i^{th} archive sample.

2.3.1. Spatial coverage sampling

As noted in Section 2.1.1 above, our prediction domain is defined as those nodes on a 500-m raster grid across SNNPR where the probability of cropping activity exceeds 0.3. These are the locations at which we wish to form predictions of soil properties from analyses of sampled material from the EthioSIS archive. The sample units, in contrast, are items in the archive, with associated coordinates, $\{\mathbf{s}_1, \ldots, \mathbf{s}_m\}$.

The basic subsample of EthioSIS samples that we draw from the archive was to be a spatial coverage sample (De Gruijter et al., 2006). A spatial coverage sample gives good coverage of a region of interest in the sense that the average distance from a random location in the region to the nearest sample point is minimized. In a uniform infinite plane a spatial coverage sample can be achieved on a grid of sample points at the vertices of tesselating equilateral triangles, but in finite irregular regions simple rules cannot be applied. One useful numerical approach is to take a regular discretization of the region (such as our 500-m grid nodes) and to partition these into n clusters formed by a k-means algorithm on their rectilinear coordinates. This was implemented by Walvoort et al. (2010) in the spcosa package for the R platform. Our approach was to use k-means clustering on the rectilinear (UTM 37 N) coordinates of the nodes in the prediction domain to form n clusters for some specified n. If $\mathscr{D}$ denotes the set of n cluster centroids $\mathbf{c}_i, i = 1, \ldots, n$ created this way (each a vector of UTM37N coordinates), where $\mathbf{c}_i \in \mathscr{D}$ denotes the i^{th} centroid, then our objective is to find the subset of archive locations $s \subset \mathscr{S}$ such that for each $\mathbf{c}_i \in \mathscr{D}$ there exists some $\mathbf{s}_j \in s$ such that the Euclidean distance

$$\left|\mathbf{c}_i - \mathbf{s}_j\right| = \min\{|\mathbf{c}_i - \mathbf{s}_k|\}_{k \in [1,m]}.$$

In general the number of elements in s, $|s| = n$, but it is possible that $|s| < n$, implying that one or more $\mathbf{s}_i \in s$ is the nearest neighbouring sample point in the archive to more than one cluster centroid. For all cases in this study there was one unique nearest neighbour to each cluster centroid. We illustrate this in Fig. 5 which shows a notional extract from the prediction domain, with the nodes of the domain shown as grey symbols. The centroids of nine clusters are shown (large open circles), and notional EthioSIS sample locations are shown as black symbols, the nearest EthioSIS site to each of the centroids is shown as a black triangle, these are the EthioSIS sites which would be selected for a spatial coverage subsample, (other EthioSIS sites are shown as black discs).

Fig. 6a shows the cluster centroids for $n = 180$ across SNNPR. The black triangles in Fig. 6b correspond to the locations of the nearest-neighbouring sample points in the EthioSIS archive, selected to form the spatial coverage subsample.

2.3.2. Supplementary points

The spatial coverage sample from the archive, $\mathbf{s}_j \in s \subset \mathscr{S}$, provide a sound basis for spatial prediction by interpolation, using the empirical best linear unbiased predictor (E-BLUP) for an appropriate spatial mixed model. In this case here our linear mixed model has an unknown mean as the only fixed effect, and the E-BLUP is the ordinary kriging prediction (Stein, 1999). However, the spatial coverage sample is not, in general, sufficient to support the estimation of the linear mixed model itself because observations at short distances, relative to the spacing of the coverage sample, are needed to estimate variance parameters. This was explored by Lark and Marchant (2018), who suggested that about 10% of the total sample size could be selected as close-paired observations to spatial coverage points.

We considered subsamples of the EthioSIS archive consisting of a total of N sample locations, comprising $M = N/10$ close-pair samples aligned with M of the $n = N - M$ spatial coverage points. After selection of the spatial coverage sample, as described in the previous section, M of these were selected with the lcube function from the BalancedSampling package for R (Grafström and Lisic, 2019). This uses the cube method of Deville and Tillé (2004) to select a sub-sample which is spatially

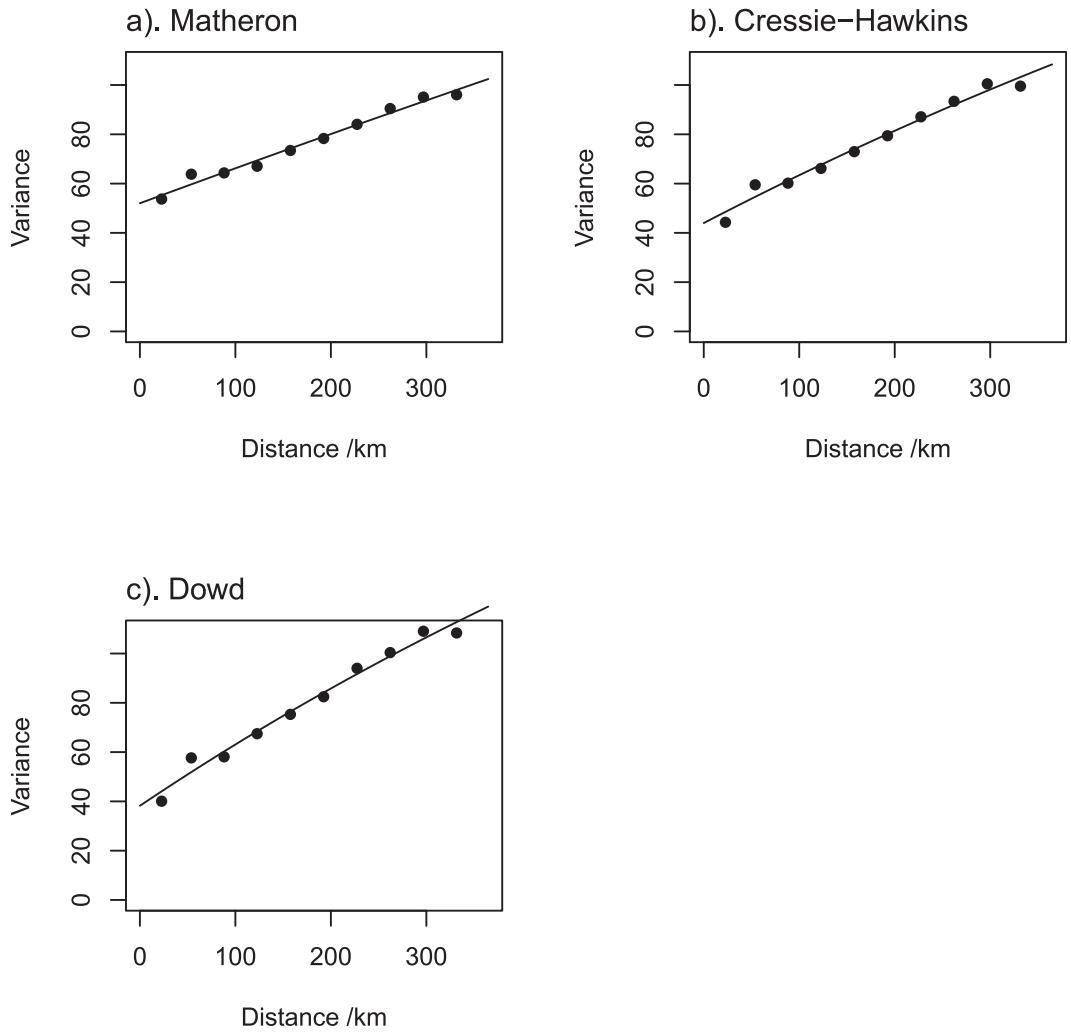

Fig. 3. Empirical variograms of soluble Se concentration (transformed to natural logarithms) forGeoNutrition soil data at or south of latitude 10° north. The variogram estimators are those of a) Matheron (1963), b) Cressie and Hawkins (1980) and c) Dowd (1984). The continuous lines are exponential variogram functions fitted by weighted least squares.

balanced (i.e. for which the mean coordinates are close to those of the sampled set), and spatially spread. Each of the spatial coverage sample units had an equal inclusion probability for selection for a close pair.

Once the close pair sites were selected each one was considered in turn, and the nearest neighbouring observation in the EthioSIS set, $\mathscr{S}$ not already selected for sampling was identified. In Fig. 6b these close-pair sites are shown as yellow triangles.

2.3.3. *Criteria for selection of a sampling density*

With a total of N points selected from the sample archive, comprising n spatial coverage points and M close neighbours, the ordinary kriging variance for soluble S and for organic I at each node of the prediction domain grid was computed, using the variogram parameters presented in Table 2. If we denote the value of the variogram according, to the model parameters in Table 2, for the interval between the locations of the i^{th} and j^{th} observations in the subset of archive sites by $\gamma(\mathbf{s}_i, \mathbf{s}_j)$ and the value for the interval between the i^{th} archive location and the location in the prediction domain at which an ordinary kriging prediction is required, $\mathbf{c}_0$ by $\gamma(\mathbf{s}_i, \mathbf{c}_0)$, then the ordinary kriging variance is given by

$$\sigma^2_{\text{OK}} = \mathbf{b}^{\text{T}}\mathbf{A}^{-1}\mathbf{b}, \tag{1}$$

where

$$\mathbf{b} = \begin{bmatrix} \gamma(\mathbf{s}_1, \mathbf{c}_0) \\ \gamma(\mathbf{s}_2, \mathbf{c}_0) \\ \cdot \\ \cdot \\ \cdot \\ \gamma(\mathbf{s}_N, \mathbf{c}_0) \\ 1 \end{bmatrix}$$

and

$$\mathbf{A} = \begin{bmatrix} \gamma(\mathbf{s}_1, \mathbf{s}_1) & \gamma(\mathbf{s}_1, \mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_1, \mathbf{s}_N) & 1 \\ \gamma(\mathbf{s}_2, \mathbf{s}_1) & \gamma(\mathbf{s}_2, \mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_2, \mathbf{s}_N) & 1 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \gamma(\mathbf{s}_N, \mathbf{s}_1) & \gamma(\mathbf{s}_N, \mathbf{s}_2) & \cdots & \gamma(\mathbf{s}_N, \mathbf{s}_N) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

The ordinary kriging variance is the expected squared error of the prediction. It is the quantity minimized in ordinary kriging.

We considered a range of total sample sizes, N, up to 300, and added in 1000 as total sample size for comparison. In each case 10% of the samples were allocated as close-paired observations to a spatial coverage sample comprising the remaining observations. The locations of the spatial coverage sample points and close pairs were selected as

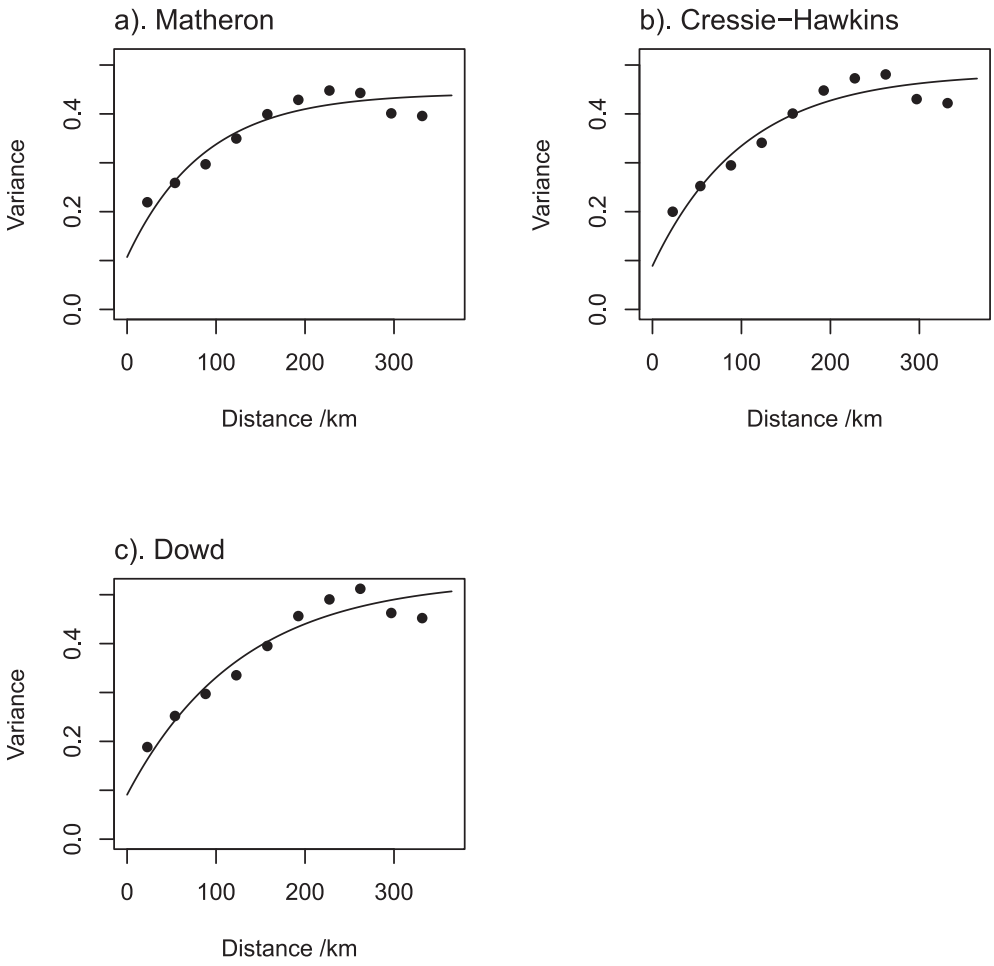

Fig. 4. Empirical variograms of organic I concentration for GeoNutrition soil data at or south of latitude 10° north. The variogram estimators are those of a) Matheron (1963), b) Cressie and Hawkins (1980) and c) Dowd (1984). The continuous lines are exponential variogram functions fitted by weighted least squares.

Table 2
Variogram parameters for selected models, with cross-validation outputs, for natural log of soluble Se and organic I concentration in GeoNutrition soil samples at or south of latitude 10° north.

Variable	Estimator	c_0	c_1	ϕ	Median SSPE	95% confidence interval for SSPE with valid model
log Se_{sol}	Matheron	0.107	0.336	86.1	0.44	[0.38,0.53]
I_{org}	Matheron	52	716	5000	0.40	[0.38,0.53]

described in Section 2.3 above. We computed, for each variable, the ordinary kriging variance at each location in the prediction domain.

The ordinary kriging variance varies in space due to the variation in the numbers and locations of archive observations. It is smallest near to an archive location selected for use, and largest at locations furthest from such observations. As a summary we extracted the 90th percentile of the kriging variance across the whole prediction domain for each variable, and plotted this against the total sample size. The results are shown in Fig. 7 (a) for log-transformed soluble Se and (b) for organic I.

It was decided that a subsample of 200 would be satisfactory. This was based on the computed ordinary kriging variance, and the shape of the plots in Fig. 7, which suggested that the marginal reduction in kriging variance for further increases in the total sample size would be small.

The sample locations for the total sample of 200 were extracted and provided to colleagues for extraction from the EthioSIS archive.

2.4. Subsampling the archive

2.4.1. Procedure for subsampling

All soil samples archived through the EthioSIS were stored at Ethiopia's National Soil Testing Center. Each sample has a unique identification number and information on geographic location (the easting and northing (WGS system) and elevation in meter). Lab technicians used the unique sample Identification Numbers (IDs) to identify the soil samples located in the 200 pre-identified locations. Once identified, 50 mg of soil was extracted from the each archived sample and placed in a plastic bag which was then sealed to prevent any cross-contamination.

2.4.2. Problems encountered

Of the 200 archive samples identified for extraction, 36 could not be found, presumed lost. For each of the missing locations the three nearest neighbours, excluding any in the original subset of 200, were identified, and the IDs provided to the colleagues doing the extraction with the

closest (first-choice substitute) and furthest of the three indicated in each case. By selection from one of these candidate samples a replacement samples was found for half (18) of the missing samples, but none of the three were available in the archive for the remaining 18 sites. Because of limited time to work in the archive a total sample size of 182 was accepted.

3. Planned analyses and criteria for evaluation

The sub-samples were then shipped to the University of Nottingham for analyses of organic iodine and soluble selenium concentrations as described above.The following subsections describe the procedures which will be followed when these analyses are complete.

3.1. Initial data screening

In each extraction batch we shall include three operational blanks (one for each fraction) to check for potential contamination. Batches will be rejected if blanks show apparent contamination (i.e., 3 times greater than the background signal of the ICP-MS). The reproducibility of the analysis is tested by repeating at least 10% of the samples in duplicate which allows us to test the variation within the analytical run and between runs. The precision is determined by calculating the relative standard deviation (RSD; %) which is the ratio of the within-duplicate standard deviation to the overall standard deviation of the material expressed as a percentage. Values of the RSD are expected to be $< 10\%$. When we suspect, based on our knowledge, that the elemental concentration of a sample is an outlier, we further investigate that sample by re-analysing in triplicates.

3.2. Exploratory data analysis

We shall compute the same summary statistics and plots of the SNNPR data as presented in this paper for the GeoNutrition data. If any observations appear to be outliers, then we shall discuss these with EB and AM to see whether there are any technical grounds for exclusion, although this is unlikely given the initial screening. If outliers remain, then we shall complete the analyses below with these included, and repeat the same analysis after exclusion of the outliers.

3.3. Geostatistical analysis

3.3.1. Primary analyses

By primary analyses we mean the analyses to test our central proposition that the variogram estimated from the subset of GeoNutrition data south of 10 degrees latitude provides a basis to select a subset of the EthioSIS soil physical archive specimens for spatial prediction of soluble Se and organic I across SNNPR by ordinary kriging. This will be deemed to have been validated if the ordinary kriging variances from the subset show comparable dependence on sample density and comparable absolute values to those displayed in Fig. 7. If the spatial dependence of the new data from the SNNPR EthioSIS subset is appreciably different from the GeoNutrition data used here, then the kriging variance might decline more rapidly, or more slowly, as a function of sample density than is shown in Fig. 7, and the absolute values of the kriging variance may be different. Our hypothesis is that the 90th percentile of the kriging variances calculated at the nodes of the prediction domain from the variogram models estimated from the new SNNPR EthioSIS analyses will vary by no more than 10% from the values expected from the variogram computed on the GeoNutrition data, and the location of the selected EthioSIS sites (noting that there were eventually fewer of these than orginally planned due to missing material).

We shall use the same procedures described above to estimate and cross-validate variograms for the variables on the SNNPR data. When this is done, we shall compute the kriging variances across the same nodes of the prediction domain. We shall plot the percentiles of these kriging variances against the corresponding percentiles based on the variogram parameters in Table 2 for the GeoNutrition data. This will show whether the GeoNutrition variogram tends to over- or under-predict the kriging variances obtained with the data from the SNNPR subset.

Because the absolute values of the SNNPR and GeoNutrition variances may differ, we shall compute the kriging variances for the same subsets of SNNPR EthioSIS locations used to produce Fig. 7(a) and (b), and standardize these by the *a priori* variance (i.e. the sill variance) of the corresponding variograms. This will show whether (i) the magnitude of the kriging variance (90th percentile) for the GeoNutrition variogram, and that for the newly-estimated SNNPR variograms differ as a proportion of the variance of a notional independent sample and (ii) whether the reduction in kriging variance (90th percentile), as a proportion of the a prior variance, on changing the total sample size from 150 to 200 to 250 differs for the two variograms. This latter analysis will show whether there are practically significant differences in the scale-dependence of the target variables between the SNNPR and GeoNutrition data sets.

3.3.2. Secondary analyses

In the secondary analysis we shall follow the procedures of Lark et al. (2017) to fit a linear mixed model to the new data on soluble Se and organic I by a Bayesian method. We shall use the same code for this analysis, but shall adjust the range of the uniform prior for the variance terms to $[0, 10\nu]$ for each variable where ν is the sample variance.

The Bayesian analysis will provide empirical posterior probability densities for the variogram parameters for each variable. We shall compare these with the estimated parameters from the GeoNutrition data set in Table 1 of this paper. We shall also use them to compute the probability density functions for the 90th decile of kriging variance across the prediction domain for alternative subsets from the SNNPR archive, and compare this with the values shown in Fig. 7.

4. Discussion

This is a pre-registered experiment, so final results will be presented subsequently. The activities reported above highlight a number of questions which require attention in further work on the problem of how best to subsample a soil archive in the light of our eventual results.

The first question relates to the assumptions made about the spatial dependence for a variable not yet analysed. In this study we have used data on organic I and soluble Se from measurements made in adjacent regions of Ethiopia. However, this is not without its problems. In this study we were able to show, using the soil map units of the Soil Atlas of Africa, that there were clear similarities in the general soil conditions of the SNNPR region and the adjacent reference regions for which we had data. However, we found a pronounced north–south trend in the reference observations, which meant that a more-or-less arbitrary decision had to be made on how to identify a subset of the reference data to be used to obtain a variogram model to use for the subsampling decisions. It might be possible to identify appropriate reference data more objectively, perhaps if these and the EthioSIS had first been analysed by laboratory-based spectral methods. One might identify a subset of the reference data which fall within some envelope of spectral values which includes the data from the archive to be subsampled.

Another issue arises from differences in the sample support between the GeoNutrition survey and the EthioSIS survey. First, there could be systematic differences between the two sets of measurements because the former were from depth interval 0–15 cm whilst the latter was from interval 0–20 cm. For our purposes this difference matters only in so far as it affects the variance of the observations. We hypothesize that the effect is small, but may be relatively larger for organic I than for soluble Se because the latter might be expected to show marked variation with depth reflecting the distribution of organic matter. Second, whilst both

samples are composite, the GeoNutrition samples were each a composition of five cores from within a circle of radium 5.6 m (area 100m^2) whilst the EthioSIS samples were each formed from nine aliquots, a central core and eight at locations 15 m from the centre. This difference in support might result in differences in the empirical variograms. If the GeoNutrition variograms were estimated from data each obtained from a single core, then they could be regularized to reflect the EthioSIS support (see Webster and Oliver, 2007). This is not the case, so there is no direct solution. Because the difference in support is relatively small, we hypothesize, again, that the effect on the variogram shape, and absolute variances will be small. This will be tested when we finally analyse the data.

An alternative approach would be not to assume that a reference data set could be treated as a proxy for the archive soil values, yet to be determined. Rather, one might subsample the archive in more than one phase, completing the laboratory analyses at the end of each phase and then analysing these data statistically to develop a spatial model used in the selection of the next subset of archive material. Marchant and Lark (2006) proposed such an adaptive sampling strategy for field soil sampling. Early phases of sampling are focused on development of the spatial model, and then the final phase uses this model to select a final set of sample points to complete the sample for the final analysis to address the project objective. The feasibility of this approach will depend on the time available, and whether the archive and laboratory are in reasonably close proximity.

For some variables, systematic examination of the literature might provide a plausible variogram model to use in particular conditions. See, for example, the average variograms published by Paterson et al. (2018). This could be used either directly to compute kriging variances for proposed subsets of archive samples, or to provide a prior variogram model for the adaptive approach described above. However, no such information was available for our particular variables.

Second, we note that a not insignificant number of our initial set of selected archive specimens were not found. This underlines the importance of good record keeping for a physical archive. We propose that such an archive should have a catalogue, in which one may record missing samples, samples for which material has been lost or used in new analyses, and samples which further analysis has suggested might be contaminated or otherwise atypical. This would be useful in further use of the archive so that scientists have a definitive list of available material. Field notes from the original sampling campaign could also usefully be linked to such a catalogue.

Third, we note that our resampling of the EthioSIS archive was done specifically to allow spatial mapping of the variables to be analyses by ordinary kriging. For this reason the criterion for selection was the ordinary kriging variance. If we had other objectives then other criteria might be more appropriate. For example, if the objective were to estimate a spatial mean of the new variable across the sample region, or a few large subregions, then a spatially balanced subsample would be a good option (Deville and Tillé, 2004). If the aim were to calibrate a predictive model for the new soil variable based on spectral measurements already made on the samples, or exhaustive covariates which can be linked to all the samples (e.g. remote sensor data), then one might select a subset which is 'spread' in the feature space defined by the covariates, for example by application of the cube algorithm (Grafström and Tillé, 2012).

Finally, we make an observation about this preregistered study. On submission of this paper we wait to see the soil data, and how far our analysis has allowed us to be efficient in selection of a subset of the archive for further analysis. Our hypothesis is that the information from the reference observations in the adjacent regions will allow us to guide the selection of the subset rationally, such that the kriging variances at the nodes of the prediction domain computed from the new data match our expectations from the application of the variograms estimated from the GeoNutrition data within the tolerances given above. This hypothesis is subject to a genuine test, because no post hoc decisions about criteria for assessment, subregions of the reference data to use, strategies for outlier identification etc. can be made. We believe that such an approach is a useful contribution to making soil science more reproducible.

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

The GeoNutrition projects, under which the soil sampling in Amhara, Tigray and Oromia Regions was undertaken, were funded by the Bill & Melinda Gates Foundation [INV-009129] and the UKRI Biotechnology and Biological Sciences Research Council (BBSRC)/Global Challenges Research Fund (GCRF) [BB/P023126/1]. The funders were not involved in the study design, the collection, management, analysis, and interpretation of data.

The soil samples from the SNNP region were collected, processed, and archived by the EthioSIS, a project implemented by the Ethiopian Government's Agricultural Transformation Agency (ATA) in collaboration with the Ministry of Agriculture, Livestock, Fisheries, and Cooperatives. The project has benefited immensely from technical and financial support from a wide range of institutions and development partners.

The boundaries, denominations and any other information shown on the maps in this paper do not imply any judgment about the legal status of any territory or constitute any official endorsement or acceptance of any boundaries on the part of any Government.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.geoderma.2022.116013.

References

AfSIS, 2015. New cropland and rural settlement maps of Africa.http://africasoils.net/2015/06/07/new-cropland-and-rural-settlement-maps-of-africa.

Belay, A., Gashu, D., Joy, E.J.M., Lark, R.M., Chagumaira, C., Likoswe, B.H., Zerfu, D., Ander, E.L., Young, S.D., Bailey, E.H.B., Broadley, M.R., 2021. Zinc deficiency is highly prevalent and spatially dependent over short distances in Ethiopia. Sci. Rep. 11, 6510.

Belay, A., Joy, E.J.M., Chagumaira, C., Zerfu, D., Ander, E.L., Young, S.D., Bailey, E.H., Lark, R.M., Broadley, M.R., Gashu, D., 2020. Selenium deficiency is widespread and spatially dependent in Ethiopia. Nutrients 12, 1565.

Black, H., Bellamy, P., Creamer, R., Elston, D., Emmett, B., Frogbrook, Z., Hudson, G., Jordan, C., Lark, M., Lilly, A., Marchant, B., Plum, S., Potts, J., Reynolds, B., Thompson, P., Booth., P., 2008. Design and operation of a uk soil monitoring network. Environment Agency, Bristol, Science Report – SC060073.

Botoman, L., Munthali, M.W., Chimungu, J.G., Mossa, A.-W., Young, S.D., Bailey, E.H., Ander, E.L., Lark, R.M., Broadley, M.R., Nalivata, P.C., 2020. Increasing zinc concentration in maize grown under contrasting soil types in Malawi through agronomic biofortification: Trial protocol for a field experiment to detect small effect sizes. Plant Direct 4, e00277.

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

Cressie, N., Hawkins, D., 1980. Robust estimation of the variogram. Math. Geol. 12, 115–125.

De Gruijter, J., Brus, D., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer-Verlag, Berlin.

Deville, J.C., Tillé, Y., 2004. Efficient balanced sampling: the cube method. Biometrika 91, 893–912.

Dowd, P.A., 1984. The variogram and kriging, robust and resistant estimators. In: Verly, G. (Ed.), Geostatistics for Natural Resources Characterization. D. Reidel, Dordrecht, pp. 91–106.

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.
Gelaw, A., Bellete, T., Kasahun, B., Demiss, M., 2018. Recent developments in soil fertility mapping and fertilizer advisory services in ethiopia: A position paper. Paper presented at the National Consultative Workshop, Adama, Ethiopia, March 2018, https://www.researchgate.net/publication/327764748_Recent_Developments_in_Soil_Fertility_Mapping_and_Fertilizer_Advisory_Services_in_Ethiopia_A_Position_Paper.
Gneiting, T., 2013. Strictly and non-strictly positive definite functions on spheres. Bernoulli 19, 1327–1349.
Grafström, A., Lisic, J., 2019. BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.5. https://CRAN.R-project.org/package=BalancedSampling.
Grafström, A., Schelin, L., 2014. How to select representative samples. Scand. J. Stat. 41, 227–290.
Grafström, A., Tillé, Y., 2012. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24, 120–131.
Hijmans, R.J., 2019. geosphere: Spherical Trigonometry. R package version 1.5-10.https://CRAN.R-project.org/package=geosphere.
Jenkinson, D.S., Rayner, J.H., 1977. The turnover of soil organic matter in some of the Rothamsted classical experiments. Soil Sci. 123, 298–305.
Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Spaargaren, O., Tahar, G., Thiombiano, L., Van Ranst, E., Yemefack, M., Zougmore, R., 2013. Soil atlas of africa. European Commission. Publication Office of the European Union, Luxembourg.https://esdac.jrc.ec.europa.eu/content/soil-map-soil-atlas-africa.
Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51, 137–157.
Lark, R.M., Hamilton, E.M., Kaninga, B., Maseka, K.K., Mutondo, M., Sakala, G.M., Watts, M.J., 2017. Planning spatial sampling of the soil from an uncertain reconnaissance variogram. Soil 3, 235–244.
Lark, R.M., Knights, K.V., 2015. The implicit loss function for errors in soil information. Geoderma 251–252, 24–32.
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.
Ligowe, I.S., Phiri, F.P., Ander, E.L., Bailey, E.H., Chilimba, A.D.C., Gashu, D., Joy, E.J.M., Lark, R.M., Kabambe, V., Kalimbira, A.A., Kumssa, D.B., Nalivata, P.C., Young, S.D., Broadley, M.R., 2020. Selenium (Se) deficiency risks in sub-saharan african food systems and their geospatial linkages. Proc. Nutr. Soc. 79, 457–467.
Lowe, N.M., Zaman, M., Moran, V.H., Ohly, H., Sinclair, J., Fatima, S., Broadley, M.R., Lark, R.M., Zia, M.H., Ander, E.L., Sharp, P., Bailey, E.H., Young, S.D., Jaffar Khan, M., 2020. Biofortification of wheat with zinc for eliminating deficiency in Pakistan: Study protocol for a cluster-randomised, double-blind, controlled effectiveness study (bizifed2). BMJ Open 10, e039231.
Marchant, B.P., Lark, R.M., 2006. Adaptive sampling for reconnaissance surveys for geostatistical mapping of the soil. Eur. J. Soil Sci. 57, 831–845.
Matheron, G., 1962. Traité de Géostatistique Appliqué, Tome 1. Memoires due Bureau de Recherches Géologiques et Minières, Paris.
Mathers, A.W., Young, S.D., McGrath, S.P., Zhao, F.J., Crout, N.M.J., Bailey, E.H., 2017. Determining the fate of selenium in wheat biofortification: an isotopically labelled field trial study. Plant Soil 420, 61–77.
McBratney, A.B., Webster, R., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables – II: Program and examples. Comput. Geosci. 7, 255–365.
Paterson, S., McBratney, A.B., Minasny, B., Pringle, M.J., 2018. Variograms of soil properties for agricultural and environmental application. In: McBratney, A., Minasny, B., Stockmann, U. (Eds.), Pedometrics. Springer, pp. 623–667.
R Core Team, 2020. R: A language and environment for statistical computing. 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 Manag. 21, 353–362.
Rawlins, B.G., McGrath, S., Scheib, A., Breward, N., Cave, M., Lister, T., Ingham, M., Gowing, C., Carter, S., 2012. The advanced soil geochemical atlas of England and Wales. British Geological Survey, Nottingham, UK.
Shetaya, W.H., Young, S.D., Watts, M.J., Ander, E.L., Bailey, E.H., 2012. Iodine dynamics in soils. Geochim. Cosmochim. Acta 77, 457–473.
Stein, M.L., 1999. Interpolation of spatial data: some theory for kriging. Springer, New York.
Walsh, M., Wu, W., Simbila, W.J., Levy, M.A., Borkovska, O., Schmidt, J., 2019. Geosurvey data prediction workflows. OSF. https://doi.org/10.17605/OSF.IO/VXC97.
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. John Wiley & Sons, Chichester.
World Reference Base, 2007. World reference base for soil resources. FAO, Rome.