Performance of linear mixed models and random forests for spatial prediction of Soil pH.

Mirriam Makungwe[1], Lydia Mumbi Chabala[1], Benson H Chishala[1], R Murray Lark[2]

[1]Department of Soil Science, University of Zambia, School of Agricultural Sciences, P.O. Box 32379, Lusaka, Zambia.

[2]School of Biosciences, University of Nottingham, Sutton Bonington, Nottinghamshire, LE12 5RD, UK.

Email address:

mirriammakungwe.tolopu@gmail.com (M. Makungwe), lchabala@unza.zm (L. M. Chabala), bchishala@unza.zm (B. H. Chishala), Murray.Lark@nottingham.ac.uk (R Murray Lark).

Author notes

Correspondence concerning this article should be addressed to Mirriam Makungwe, Department of Soil Science, School of Agricultural Sciences, University of Zambia P.O. Box 32379, Lusaka, Zambia. E-mail: mirriammakungwe.tolopu@gmail.com

ABSTRACT

Digital soil maps describe the spatial variation of soil and provide important information on spatial variation of soil properties which provides policy makers with a synoptic view of the state of the soil. This paper presents a study to tackle the task of how to map the spatial variation of soil pH across Zambia. This was part of a project to assess suitability for rice production across the country. Legacy data on the target variable were available along with additional exhaustive environmental covariates as potential predictor variables. We had the option of undertaking spatial prediction by geostatistical or machine learning methods. We set out to compare the approaches from the selection of predictor variables through to model validation, and to test the predictors on a set of validation observations. We also addressed the problem of how to robustly validate models from legacy data when these have, as is often the case, a strongly clustered spatial distribution. The validation statistics results showed that the empirical best linear unbiased predictor (EBLUP) with the only fixed effect a constant mean (ordinary kriging) performed better than the other methods. Random forests had the largest model-based estimates of the expected squared errors. We also noticed that the random forest algorithm was prone to select as "important" spatially correlated random variables which we had simulated.

Keyword: Linear Mixed Models; REML-EBLUP; Random forests, Spatial prediction of soil pH.

1. INTRODUCTION

Soil maps describe the spatial variation of soil types and provide important information on spatial variation of soil properties (Kempen et al., 2010). Mapping of soil properties is important as it provides policy makers with a synoptic view of the state of the soil, and agricultural stakeholders with information about where soil problems might occur (Lark et al., 2019). Soil maps are generated using various soil mapping methods which can be divided into conventional and pedometric approaches (Kienast-Brown et al., 2010; Hengl, 2003).

Conventional soil survey represents soil variation in terms of profile classes and corresponding map legend units. It can provide a basis for spatial prediction of soil properties and may also serve as a structure for recording substantial information on soil management and for systematizing knowledge of the distribution of soils in the landscape. Conventional approaches were based largely on manual processes which are costly and time consuming (Kienast-Brown et al., 2010) mainly because of long fieldwork periods (Moonjun et al., 2010). Pedometric approaches are based on the application of mathematical and statistical methods for the primary purpose of predicting the values of soil properties where these have not been observed directly (McBratney et al., 2000). A well-established statistical approach to doing this is the application of model-based geostatistics (Stein, 1999; Diggle and Ribeiro, 2007). In this approach the variation of the soil is represented in a linear mixed model (LMM) as a combination of fixed effects (which may be a constant unknown mean, or a function of predictive covariates such as remote sensor data), and random effects, including Gaussian random fields which exhibit spatial correlation. The parameters of the LMM model can be estimated by Residual Maximum Likelihood (REML) method developed by Patterson and Thompson (1971), which allows parameters of the random effects to be estimated with small bias arising from uncertainty in the fixed effects (Kitanidis, 1987; Swallow and Monahan 1984; Zimmerman and Zimmerman, 1991; Lark and Cullis, 2004) . When the model is fixed, values of the soil property at unsampled sites can be obtained by the empirical best linear unbiased predictor (EBLUP) (Stein, 1999; Lark et al., 2006; Lark and Webster, 2006; Minsay and McBratney, 2007).

There has been a growing interest in the potential of machine learning methods (e.g. Breiman, 2001) as an alternative to statistical modelling for spatial prediction of soil properties (Hengl et al., 2015; Behrens and Scholten, 2007). The main difference between geostatistical approaches and random forest is that geostatistics is based on a statistical model. This provides a basis for formal inference about the validity of the model (including the task of selecting which covariates to use in prediction), and for producing a prediction distribution at unsampled sites of interest. One may then derive point predictions from this distribution (typically the mean), and measures of uncertainty. On the other hand, machine learning methods such as random forests, are predictive tools applied to identify empirical relationships between the target variable in a training data set and associated predictive covariates and to extrapolate these to unsampled sites. With no model there can be no formal inference, but empirical approaches, based on internal cross-validation are used, for example, to evaluate the evidence that a particular variable is predictive. One particular strength of the geostatistical approach is that the estimation of coefficients for predictor variables, and inferences about them, are based on a model of the spatial dependence of the random variation. This accounts for the fact that data which are strongly spatially clustered are likely to be correlated, and so do not provide independent evidence to support the fitted model.

In the study reported here our objective was to assess approaches for digital mapping of soil pH at national scale across Zambia to support evaluation of land potential for rice production. Legacy data on soil pH were available from a previous national survey. As with many such surveys, this followed a two-stage design, and so the observations were spatially clustered. In addition we had access to various exhaustive environmental covariates which could be regarded as potential predictor variables for soil pH.. We compared different forms of linear mixed model, and prediction with the random forest using a validation subset of the data. Prediction errors were evaluated at the validation locations by comparing predictions with observed values. The selection of the validation subset, and the quantification of the uncertainty from the observed prediction errors had to take account of the spatial clustering of the observations in the legacy data. Because of this clustering, no subset could be regarded as independent random observations.

2. THEORY

2.1 Linear Mixed Model

The theory of residual maximum likelihood (REML) in combination with the empirical best linear unbiased predictor (EBLUP) for spatial interpolation has been illustrated and described in detail by Lark et al., 2006. The LMM takes the form

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

where $\mathbf{z}$ is a set of observations of the random variable at sampled locations, $\mathbf{M}$ is the design matrix of fixed effects, which could include covariates such as topographic attributes, $\boldsymbol{\beta}$ is the vector of the fixed effects parameters or regression coefficients, $\mathbf{S}$ is the design matrix of random effects (which is an identity matrix unless analytical duplicate observations are included), $\boldsymbol{\eta}$ is a random effect, a Gaussian random variable which has a mean of zero and, in the spatial setting, a covariance matrix which expresses spatial dependence, $\boldsymbol{\varepsilon}$ is an independently and identically distributed Gaussian residual of mean zero and variance σ^2. These two random components have a joint distribution

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

where **I** is the identity matrix and **G** is the correlation matrix of the random effect $\boldsymbol{\eta}$. Element [*i*,*j*] of **G**, at locations $\mathbf{x}_i$ and $\mathbf{x}_j$ depends only on the interval in space between them under an assumption of second-order stationarity. The lag vector $\mathbf{x}_i - \mathbf{x}_j$, under the assumption of isotropy, depends only on the scalar part of this vector, the lag distance and so

$$\mathbf{G}[i,j] = \rho\left(\left|\mathbf{x}_i - \mathbf{x}_j\right|; \alpha\right), \tag{3}$$

where $\rho(h;\alpha)$ is a correlation function of lag distance h with spatial parameters α which control how the correlation decreases with increasing distance. The term ξ is the ratio of the variance of the random effect $\boldsymbol{\eta}$ to σ^2, the variance of the residual term.

The residuals depend on the fixed effects parameters $\boldsymbol{\beta}$ in the model, and in ordinary maximum likelihood estimation the uncertainty in the estimates of the fixed effects parameters biases the estimates of the random effects parameters. To avoid this, we use residual maximum likelihood (REML) which is based on the principle that a new random variable, independent of the fixed effects, is computed by projecting the original data $\mathbf{z}$ into a residual space where the fixed effects can be filtered out (Chai et al., 2008). The log likelihood of the new random variable which we now call the residual log-likelihood because its independent of fixed effects can be expressed as;

$$\ell_R(\sigma^2,\xi,\alpha|\mathbf{z}) = -\frac{1}{2}\{\log|\mathbf{H}| + \log|\mathbf{M}^{\mathrm{T}}\mathbf{H}\mathbf{M}| + (n-p)\sigma^2 + \frac{1}{\sigma^2}\mathbf{z}^{\mathrm{T}}(\mathbf{I}-\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^{\mathrm{T}})\mathbf{z}, \quad (4)$$

Where $\mathbf{H} = \xi\mathbf{M}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}$, $\mathbf{W} = [\mathbf{M},\mathbf{S}]$ and $\mathbf{C} = \begin{bmatrix}\mathbf{M}^{\mathrm{T}}\mathbf{M} & \mathbf{M}^{\mathrm{T}}\mathbf{S} \\ \mathbf{S}^{\mathrm{T}}\mathbf{M} & \mathbf{S}^{\mathrm{T}}\mathbf{S} + \xi^{-1}\mathbf{G}^{-1}\end{bmatrix}$.

Once the covariance parameters σ^2, ξ, α have been estimated by REML, they are used to compute the estimated covariance matrix at sampled points. With the estimated covariance matrix computed, the estimated fixed effects parameter, $\widehat{\boldsymbol{\beta}}$, and predicted random effects, $\widetilde{\boldsymbol{\eta}}$, are then computed by solution of mixed model equation:

$$\mathbf{C}\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \widetilde{\boldsymbol{\eta}}\end{bmatrix} = \begin{bmatrix}\mathbf{M}^{\mathrm{T}}\mathbf{z} \\ \mathbf{S}^{\mathrm{T}}\mathbf{z}\end{bmatrix} \quad (5)$$

With the covariance matrix for the error of the estimates being:

$$\mathrm{Cov}\begin{bmatrix}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\end{bmatrix} = \sigma^2\mathbf{C}^{-1}. \quad (6)$$

Now that the covariance matrix and the fixed effects parameters have been estimated, they are used in EBLUP to predict the soil property , $\tilde{z}_p$, at unsampled locations:

$$\tilde{z}_p = \mathbf{M}_\mathrm{p}^\mathrm{T}\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\eta}}_\mathbf{p} = \mathbf{M}_\mathrm{p}^\mathrm{T}\hat{\boldsymbol{\beta}} + \mathbf{g}_\mathrm{o,p}^\mathrm{T}\mathbf{G}^{-1}\tilde{\boldsymbol{\eta}}, \qquad (7)$$

where $\mathbf{M}_\mathrm{p}$ is the design matrix for the prediction sites, $\mathbf{g}_\mathrm{o,p}$ is a vector computed from the covariance matrix of $\boldsymbol{\eta}$ with the $\boldsymbol{\eta}_\mathrm{p}$ values at the unsampled locations ($\mathrm{Cov}[\boldsymbol{\eta}, \boldsymbol{\eta}_\mathrm{p}] = \xi\sigma^2\mathbf{g}_\mathrm{o,p}$). The variance of the prediction errors, $Var[\tilde{z}_p - z_p]$, which accounts for the uncertainty in predicting the fixed effects and uncertainty in predicting the random effects is expressed as:

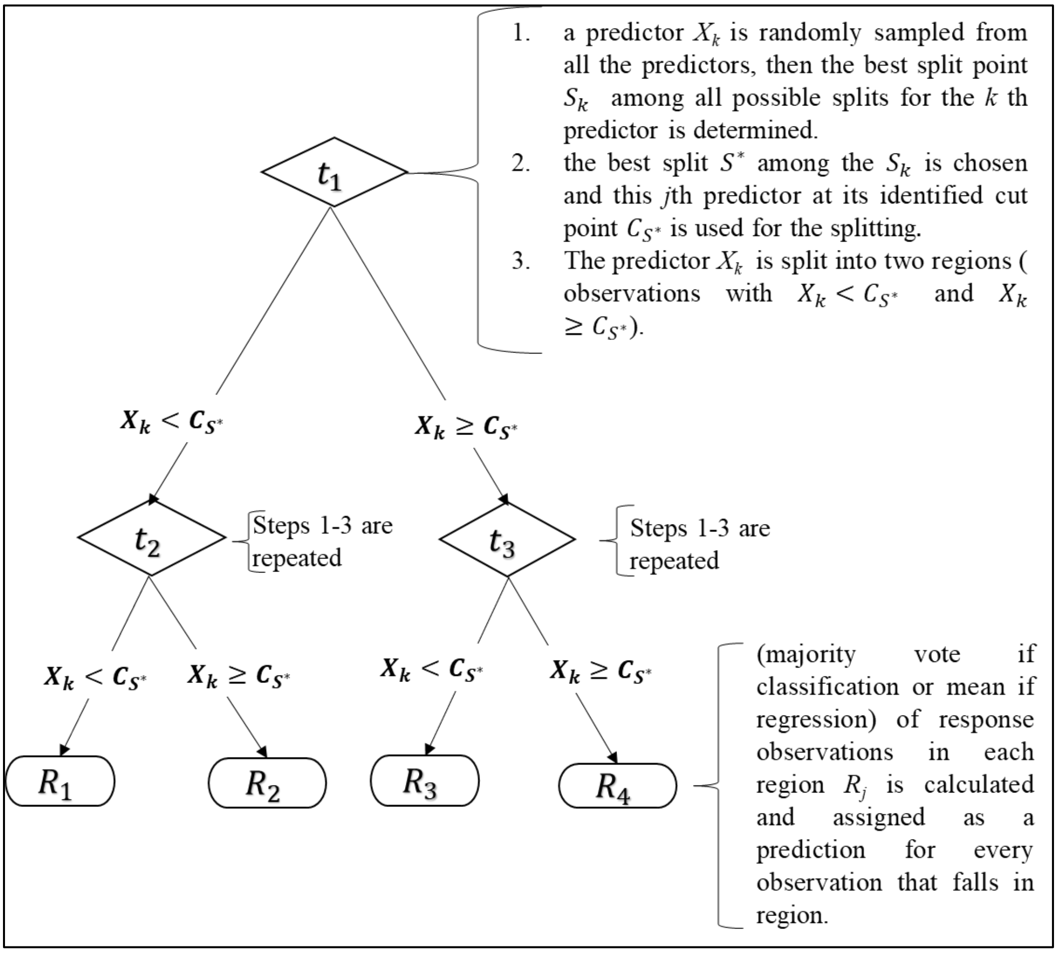
$$\mathrm{Var}[\tilde{z}_p - z_p] \quad = \sigma^2\left\{\left[\mathbf{M}_\mathrm{p}, \mathbf{g}_\mathrm{o.p}^\mathrm{T}\mathbf{G}^{-1}\right]^\mathrm{T}\mathbf{C}^{-1}\left[\mathbf{M}_\mathrm{p}, \mathbf{g}_\mathrm{o.p}^\mathrm{T}\mathbf{G}^{-1}\right] + \xi\left(\mathrm{g}_\mathrm{p,p} - \mathbf{g}_\mathrm{o.p}^\mathrm{T}\mathbf{G}^{-1}\mathbf{g}_\mathrm{o,p}\right) + 1\right\}. \qquad (8)$$

There are many variables that researchers can use as fixed effects in linear mixed models for spatial prediction of soil properties. However, it is unwise to include variables without regard for evidence that they are of predictive value, the inclusion of predictors unrelated to the target variable may inflate the prediction error variance. To avoid this, variable selection is an important step in model development. One approach to the problem is to base the inclusion or rejection of a predictor based on a hypothesis test in the LMM framework (e.g. by a log-likelihood ratio test) (Verbeke and Molenberghs, 2000). To reduce the risk of including excess predictors because of multiple hypothesis testing, one may use false discovery rate control (Lark, 2017). The false discovery rate (FDR) is the probability that a null hypothesis is true, given that it has been rejected. False discovery rate control can reduce the power to detect real predictors, and Lark, (2017) demonstrated how this problem can be reduced, while maintaining FDR control, by the method of alpha investment (Foster and Stine, 2008). This entails an initial ordering of the predictors starting with the one which, a priori (and not based on inspection of the data) is thought most likely to be of predictive value and adding in predictors in declining order of expected predictive power. In this approach the power to detect a predictor is increased by the rejection of null hypotheses early in the sequence while maintaining control of FDR. This approach has been used elsewhere for spatial prediction (Gashu et al., 2020).

A disadvantage of the LMM approach is that it assumes that the fixed effects are linear in the parameters. Such a model can represent complex and non-linear relations between soil properties and predictors, for example through the use of polynomial terms in the predictor variables, or spline basis functions, but there has been increasing interest in more flexible methods to predict soil properties from covariates, in particular the machine learning method known as the random forest.

2.2 Random Forests

The random forest is an ensemble tree-based method that combines multiple decision trees (classification or regression) to give a prediction (Breiman, 2001). A decision tree is an algorithm that involves recursive partitioning of data into several simple regions using a series of splitting rules. It is called a decision tree because these series of splitting rules can be summarised into an upside-down tree structure as illustrated in Figure 1. Figure 1 shows a structure made up of predictors ($X_1, X_2, \dots \dots, X_k$) which are split into J distinct and non-overlapping regions ($R_1, R_2, \dots, R_j$) at test node t, and the mean of the response values for the training observations in each region R_j is calculated and assigned as a prediction for every observation that falls in region R_j (James et al., 2013). When growing a decision tree, the following steps are taken; (1) at each test node t, a predictor X_k is randomly sampled from all the predictors, then the best split point S_k among all possible splits for the k th predictor is determined; (2) the best split S^* among the S_k is chosen and this jth predictor at its identified cut point C_{S^*} is used for the splitting at test node t. (3) The predictor X_k is split into two regions (observations with $X_k < C_{S^*}$ and $X_k \geq C_{S^*}$) at test node t. Steps 1-3 are repeated on all descendant nodes to grow a tree $\hat{f}(x)$ (Archer and Kimes, 2008; James et al., 2013).



1. a predictor X_k is randomly sampled from all the predictors, then the best split point S_k among all possible splits for the k th predictor is determined.
2. the best split S^* among the S_k is chosen and this jth predictor at its identified cut point C_{S^*} is used for the splitting.
3. The predictor X_k is split into two regions (observations with $X_k < C_{S^*}$ and $X_k \geq C_{S^*}$).

Steps 1-3 are repeated

Steps 1-3 are repeated

(majority vote if classification or mean if regression) of response observations in each region R_j is calculated and assigned as a prediction for every observation that falls in region.

Figure 1: illustration of a decision tree

One major limitation with decision trees is that using only one tree for prediction, results in highly unstable predictions, a small change in the data can result into a large change in the final estimated tree. To improve the performance of decision trees, Breiman (1996) introduced an algorithm called Bagging, also known as bootstrap aggregation which takes repeated (bootstrap) samples (where *B* is the number of bootstrap samples) from training set with replacement and builds a total of *B* trees $(\hat{f}^{1}(x), \hat{f}^{2}(x), \dots\dots \hat{f}^{b}(x))$ which the average of all the prediction trees $\hat{f}_{bag}(x)$ is calculated:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x) \tag{9}$$

One disadvantage of bagging is that a single predictor may dominate all trees in the bag, meaning that their outputs are strongly correlated. As a result of this the reduction in variance from the use of multiple trees is very limited (James et al., 2013). To address this, Breiman (2001) developed a random forest algorithm which is an improvement of bagging. Like bagging, random forest also takes repeated (bootstrap) samples from the training data and builds B decision trees. But in the case of random forest, when building these trees, to avoid using one strong predictor for all bagged trees, at every test node t, when splitting, every bagged tree is forced to consider only a random subset of predictors by randomly sampling a fresh m predictors from a set of k predictors, and the split is only allowed to use one of these m predictors. For regression trees, the number of m predictors considered at each split is approximately the total number of predictors divided by three ($m \approx {}^{k}/_{3}$) and for classification trees, $m \approx \sqrt{k}$. Because of this, random forest results in many uncorrelated trees which give a large reduction in variance when averaged.

The random forest algorithm has three important outputs. These are the out-of-bag Mean Squared Error (OOB Mean Squared Error), the out-of-bag R-squared and the variable importance. The RF model does not use all the data for building the tree. In each bootstrap training set, about one-third of the data are left out. The data that are left out when building the trees is called out-of-bag (OOB) data and after the trees are grown, the OOB data are used as test set to measure the strength (OOB Mean Squared Error) and correlation (OOB R-squared) of the model. In short, random forest has an inbuilt cross-validation. Variable importance is defined as the increase in prediction error when OOB data for that variable is randomly permuted while all others are left unchanged (Liaw, 2002). It analyses the contributions of each predictor to the overall results (Breiman, 2001). The algorithm randomly permutes the predictor X_m several times, breaking its original association with the response variable and asses the relevance of the predictor by using the permuted predictor together with the other unpermuted predictors to predict the response variable for the out-of-bag observations giving the difference in prediction accuracy before and after permuting. The result is a vector of importance measures for each predictor equivalent to the number of permutations. The algorithm then computes a p-value as a measure of the evidence that a variable is predictive (Strobl et al., 2007; Altmann et al., 2010). This permutation p-value is the probability of observing a permuted model (from the several number of permutations) that is equal to or better than the unpermuted model (Cummings et al., 2004).

Equation (7) presents the E-BLUP from a linear mixed model. The second term on the right-hand side corresponds to the spatial interpolation of the correlated random effect in the model. In this way the E-BLUP combines a regression-type prediction based on the predictor variables with a spatial interpolation component. As described above the random forest predicts a soil property from the predictor variables only, making no use of spatial dependence through interpolation. An attempts has been made to include spatially weighted local observations in prediction by random forests by including coordinates as predictors and using weighted buffer distances (Hengl et al., 2018), neighbouring observations and their distances to the prediction location were used by Sekulić et al., (2020). Li et al., (2011) and Viscarra Rossel et al., (2014) combined random forest with kriging, just like in regression-kriging, by calculating the random forest residuals and then kriged them to all prediction positions and then added to the results of the prediction positions. .

As described above, inferences in the random forest approach are based on an internal cross-validation procedure. This might lead to overoptimistic conclusions about a random forest model, or about the value of a particular predictor if observations from the same clusters appear in the OOB sample and in the data used to develop the trees. That is because the observations within a cluster can be expected to be strongly correlated, and so the validation of a model fitted to data on strongly correlated observations will give an unduly optimistic impression of the model's capacity to predict at an independent location.

3. CASE STUDY

3.1. Data

Soil data

This case study uses Rural Agricultural Livelihoods Survey (RALS) of 1713 soil pH data collected by Indaba Agricultural Policy Research Institute (IAPRI) in collaboration with Central Statistical Office (CSO) and Ministry of Agriculture. The sampling frame for the RALS 2012 survey was based on the 2010 Census of Housing and Population (CSO/MAL/IAPRI, 2015). Full detail of the stratified two-stage sampling design is provided by (CSO 2012). Four households were randomly selected in each Standard Enumeration Area (SEA) and soil samples were collected from the largest maize field. A composite of 10–20 sub-samples of soil collected throughout each field and each sub-sample was a composite of equal parts soil in the 0–10cm and10–20cm depth horizons. Full details on the soil collection and laboratory analysis for soil pH (determined for a soil suspension in $CaCl_2$ with a standard pH meter) are provided by Burk et al., (2019) and Chapoto et al., (2016). The spatial prediction of soil pH for Zambia using this data has been studied by Chapoto et al., (2016) who only used ordinary kriging for the prediction.

Data cleaning involved removal of spurious values in the x and y coordinates. The need for this was indicated when the raw data were first plotted, showing points lying outside the borders of Zambia. The mean coordinates of all households were computed in each Standard Enumeration Area (SEA) centroid, and then the households were removed from the data set if the notional distance to the SEA centroid exceeded 10km. This threshold value was decided after discussion with IAPRI staff about plausible values for the distance between a village in the EA and the centroid. After data cleaning, a total of 1202 soil samples were used for analysis.

The sampling pattern for the RALS survey was not designed for spatial interpolation of soil. Due to the sampling pattern, the data is strongly clustered at the level of SEAs (the SEAs are the clusters) with a total of 362 clusters. For this reason, splitting of the dataset into training set (80%) and test set (20%) was done at cluster level (the 362 clusters were split into 260 (80%) for training and 102 (20%) for validation). Figure 2 Shows the training and test clusters with the red solid dots being the training clusters and the blue solid triangles being the test clusters.

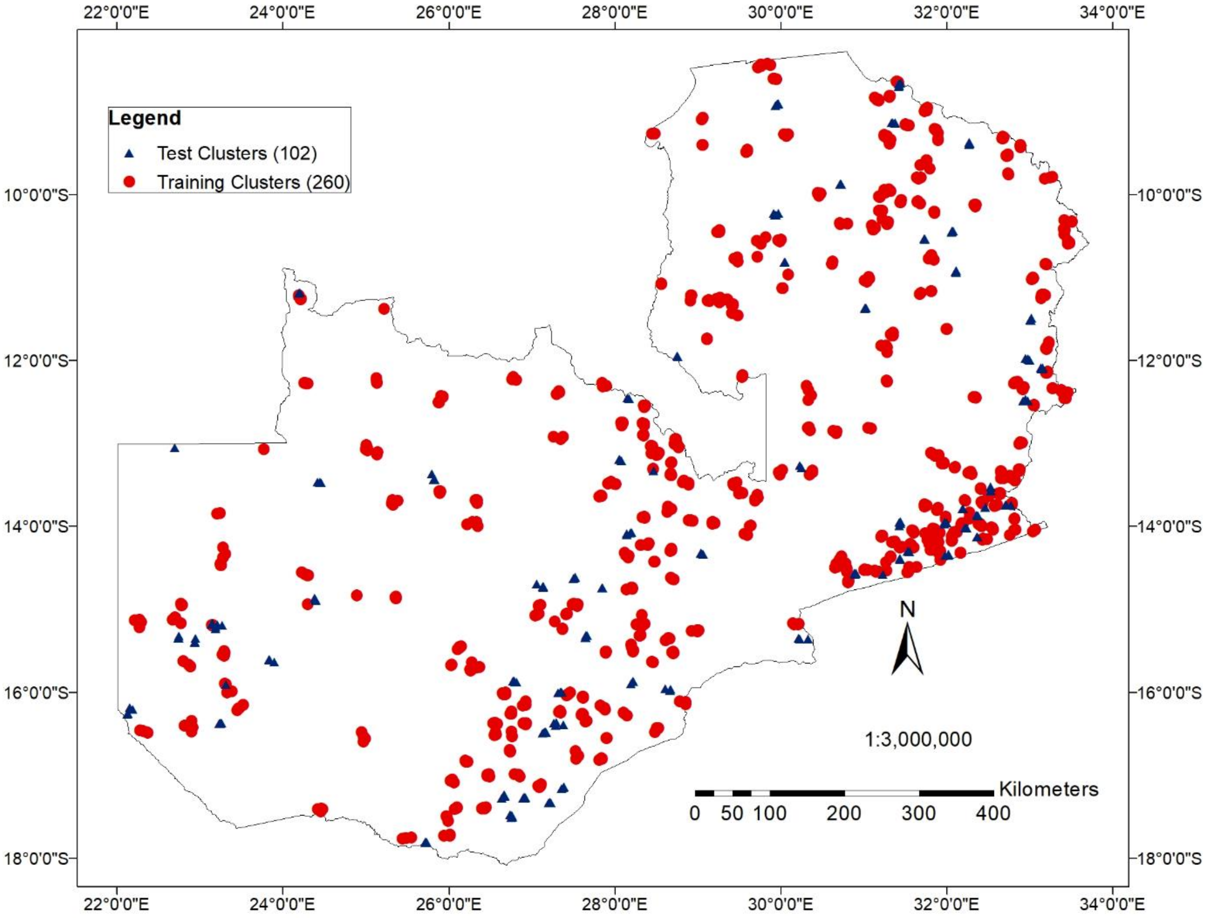


Figure 2: Cluster locations for the RALS 2012 soil data. red solid dots being the training clusters used for spatial prediction of soil pH and the blue solid triangles being the test clusters left out for validation.

Environmental Covariates

The environmental predictors available for use in this study were Soilclass, Landcover, mean annual rainfall, elevation, slope, aspect, valley depth, LS-Factor (a combination of slope and slope length, relative slope position (RSP), channel network base Level (CNBL) and Normalized difference vegetation index (NDVI).

Soilclass information was obtained from the 1:1,000,000 scale exploratory soil map of Zambia compiled by the Zambian Ministry of Agriculture, Zambia Agricultural Research Institute (ZARI) - Soil Survey Section in 1991(GRZ, 1991) and then digitized to raster format. Map units are allocated to suborders of the FAO-UNSECO classification as used in the Third Draft of the legend to the Soil Map of the World (Jahn et al, 2006)). A total of 96 soil classes were represented in the data available for model development, but these do not comprise all the classes on the map of Zambia, and so some generalization is required to develop models for spatial prediction. We, therefore, reduced the number of classes, by aggregating the classes from suborder to order level, this reduced the number of classes to 18 and all the classes in the prediction grid where represented in the training set. Land cover data for the years between 2000 and 2015 with spatial resolution of 300m were downloaded from the European Space Agency (ESA), (2017). The data presented a similar situation as that of soilclass with landcover classes in the prediction sites not being represented in the training set. We also reduced the number of landcover classes, by aggregating them as shown in Table 1.

Table 1: Aggregated landcover classes based on ESA, (2017)

New Class	ESA Class	Description
1	10	rainfed cropland
	20	irrigated or post-flooding cropland
	30	Mosaic cropland (>50%) / natural vegetation (tree
2	11	Herbaceous cover
	40	herbaceous cover) (>50%) / cropland (<50%)
	110	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)
3	12	Tree or shrub cover

	100	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)
4	50	closed to open (>15%), evergreen, broadleaved, tree cover
	60	closed to open (>15%), deciduous, broadleaved, tree cover
	61	closed (>40%), deciduous, broadleaved, tree cover
	62	open (15-40%), deciduous, broadleaved, tree cover
5	120	Shrubland
	122	Shrubland deciduous
6	130	Grassland
7	160	fresh or brackish water, flooded, tree cover
	170	saline water, flooded, tree cover
	180	fresh/saline/brackish water, flooded Shrub or herbaceous cover
8	190	Urban areas
9	200	Consolidated bare areas
	202	Unconsolidated bare areas

Mean annual rainfall data (averages from 1970 to 2000) with a spatial resolution of 1km was downloaded from WorldClim website (Fick and Hijmans, 2017). A 90-m resolution NASA Shuttle Radar Topography Mission (SRTM3) Digital elevation model (DEM) was downloaded from USGS (2019) and projected to WGS 84 UTM 35 S. The DEM was pre-processed by filling sinks using the fill sinks (Planchon/Darboux, 2001) tool in Saga GIS, and then elevation, slope, aspect, valley depth, LS-Factor (a combination of slope and slope length), relative slope position (RSP) and channel network base Level (CNBL) data was extracted from the DEM using basic terrain analysis tool in Saga GIS. MODIS land surface reflectance (MOD009GA V6) was downloaded from USGS (2019). After downloading the respective data sets, Quantum GIS was used to project the data sets to WS 84 UTM 35s and then converted to the Integrated Land and Water Information System (ILWIS) format. Then ILWIS was used to harmonise all the raster files to the same extent and cell size of 1km. Normalized difference vegetation index (NDVI) was extracted from the remote sensing images using the imageIndices of the soilassessment package for the R platform (Omuto, 2020).

3.2. Spatial Prediction of soil pH

Soil pH is an important chemical property of the soil that affects its fertility status. This is because the availability of most essential plant nutrients is influenced by the levels of pH in the soil (Jones, 2012). There are two principal processes that affect the levels of soil pH in the soil (1) the production of H^+ ions and (2) the loss of basic cations from the soil. Eleven variables were available to be considered as possible predictors for soil pH.

In section 2.1 we explained how variable selection for the LMM included false discovery rate control, to avoid over-fitting, with alpha-investment to improve the probability of retaining covariates which are predictive as predictor variables. The alpha-investment approach is most effective if the predictors can be ordered, a priori, with the one thought most likely to be predictive ranked first and so on as shown in Table 3. It must be emphasized that this ranking is based on prior considerations about the underlying process, and not on exploration of the data. We did this ordering of exhaustive environmental covariates based on how they influence the production of H^+ ions and the loss of basic cations from the soil. Rainfall was proposed as the most influential factor at national scale. Soils in environments with large annual rainfall tend to have relatively low pH due to reduced based saturation resulting from loss of basic cations by leaching (McCauley, et al., 2009; Brady and Weil, 2014). For this reason more acid soils are expected in the northern parts of Zambia (Agroecological Region Three) and in the south (Agroecological Region One) where annual rainfall is much smaller (Veldkamp et al., 1984; GRZ and UNDP, 2009). Soil class was ranked second because the soil classes represent variations in soil parent material, weathering and rejuvenation of land surfaces and development of the soils. The old, highly weathered plateau soils in the northern part of the country have lost most of the basic cations. The sandy soils in the western part are easily leached with little accumulation of basic cations. On the on the hand, the Karoo group materials in the valleys are rich in basic cations resulting in high pH values. After soil class we included topographic variables Slope, Elevation and Valley Depth. These should reflect processes such as the movement of water which carries with it dissolved basic cations from steep slope to flat areas, and the rejuvenation of weathered land surfaces which entails the removal of old highly-weathered material to reveal material with a larger content of weatherable minerals. We then included Landcover and the Normalized Difference Vegetation Index (NDVI). These will reflect effects of land management practices, including agricultural inputs, and decisions on land use which may depend on how local pH limits crop performance. The NDVI will also reflect local vigor of vegetation growth, which may be pH limited. Finally we included some further topographic variables which may reflect differences in the soil-forming environment (length-slope factor, channel network base level, relative slope position and aspect.

The data points were first projected from WGS 1984 to WSG UTM 35s. A total of 19 observations had duplicate coordinates, which were jittered by adding 100m to each of the coordinates for one site. An exploratory model was fitted to the data with all predictors included, using the likfit function of geoR package (Ribeiro and Diggle, 2001) with residual maximum likelihood (REML) as the likelihood method. The only output from this model which was examined were the residuals, for which summary statistics were calculated, and exploratory plots to check the plausibility of the assumption of normally distributed errors. In addition, the correlation model type (exponential or spherical) was identified for which the residual likelihood was largest, and this model was then used in all further analyses. During the sequence testing of hypothesis, first the null model, m_0, (with the only fixed effect a constant mean) was fitted with the likfit function and ML as the likelihood method. Then the next model, m_1, with the first predictor in the sequence was fitted in the same way. The likelihood ratio was then calculated:

$$L = 2(L_{\mathrm{m1}} - L_{\mathrm{mo}}), \qquad (10)$$

where L_{m1} is the likelihood for model m$_1$ and L_{m0} is the likelihood for the null model. If the null model is correct, then the asymptotic distribution of L is χ^2 with degrees of freedom equal to the number of additional parameters in model m$_1$ by comparison with m$_0$. If L provided evidence to reject the null model with P<0.05, then the additional predictor in model m$_1$ was retained. The second predictor in the list was then considered. When all predictors had been examined the P-values at each step were reassessed in sequence using alpha-investment as described by Lark (2017) and controlling the FDR at 0.05. Details of this approach are provided by Lark (2017), but in summary successive tests are made against a threshold P-value which depends on a quantity, the alpha-wealth, which is either augmented when null hypotheses are rejected or augmented when they are rejected. If the hypotheses are ordered so that the variables which, a priori, are expected to be good predictor variables are considered early, then this alpha investment method increases the probability of selecting predictive covariates while controlling FDR.

After variable selection, the likfit function of geoR package in R with REML as the likelihood method was used to fit two linear mixed models. One with the selected predictors as fixed effects (Kriging with an external drift) and the other with a constant mean as fixed effect (ordinary kriging). The E-BLUP prediction for both models was then calculated at the validation points.

The ranger function of ranger package (Wright and Ziegler, 2017) was used to fit the random forest model. Because random forest has an inbuilt variable selection that occurs within the model by randomly selecting variables to be used at splitting nodes, two models were fitted, one with all variables and the other with the two variables that were selected during the alpha-investment variable selection procedure. In this study, we use the ranger package in R to compute the permutation variable importance according to Altmann et al., (2010).

When predicting soil properties in space, the random forest algorithm can find apparently predictive relationships between the target variable and arbitrary spatial variables (such as digital images of human faces) when these are presented as candidate predictor variables alongside covariates which pedometricians might reasonably expect to be predictive of soil properties, (Wadoux et al., 2020). This shows that pattern recognition should not be equated to knowledge discovery. It may also suggest that the random forest algorithm is prone to overfitting, as a result of which its predictions at independent locations may be unreliable. To investigate this effect, we generated entirely random spatially autocorrelated candidate predictor variables, independent of our data, which we call null predictors. We used six spatially correlated but mutually independent null predictors, specifying a spherical variogram with a distance parameter of 100 km, nugget variance of 0 and correlated variance of 1 for each. We used the function RFsimulate from the RandomFields package for R (Schlather et al., 2015) to simulate valuses of these null predictors at the calibration locations. We then used the ranger package (Wright and Ziegler, 2017) to fit a random forest model with all predictors including the null variables as predictors and then computed the permutation variable importance according predictor p-values from the model result.

To examine the possibility of improving RF predictions by an additional kriging step (following Li et al., 2011 and Viscarra Rossel et al., 2014), residuals of the models at training points were derived (subtracting the predicted values from the observed values) and then a variogram was fitted to the residuals using likfit with a constant mean as the only fixed effect. The evidence for spatial dependence in the residuals was assessed by comparing the Akaike Information Criterion (AIC) for the fitted model and for a non-spatial alternative which are reported in likfit output.

3.3. Validation

Validation of each selected model or random forest was done using the validation data set. At each validation location the predicted soil pH was computed, and the prediction error was calculated as the difference between the predicted and observed soil pH (so a positive error is when the predicted pH exceeds the observed value). As exploratory summary statistics the mean error, median and mean square error were computed.

The validation data belong to a subset of SEA from the original survey, and as such are strongly clustered. Because of this the sample average of the squared errors may not be a good estimate of the mean square error, because the observations are not independent. A model-based approach was therefore taken to compute the expected squared error of prediction. A LMM was fitted to the prediction errors at the validation site (with a constant mean the only fixed effect). The expected square error (ESE) for each set of predictions was then computed as the sum of the squared mean error and the variances (nugget and spatially correlated) from the LMM. This is the *a priori* mean square error, i.e. the expected square error at a random location, and as such is likely to exceed the MSE computed directly from the errors of clustered data.

4. RESULTS

4.1 Variable selection

Table 2 and Figure 3 show the distribution of the residuals from the exploratory model. The histogram appears symmetrical and normal and the points on the QQ plot are close to a straight line. The residuals have octile skewness inside the range $[-0.2,0.2]$ and skewness inside $[-1,1]$, which would mean that a transformation is not normally considered necessary (Rawlins et al., 2005, Webster and Oliver, 2007).

Table 2: Statistical summary of residuals from the exploratory model

mean	Median	Variance	SD	Skewness	Octile skewness	Kurtosis
2.366e-16	-0.040	0.167	0.408	0.487	0.159	1.044

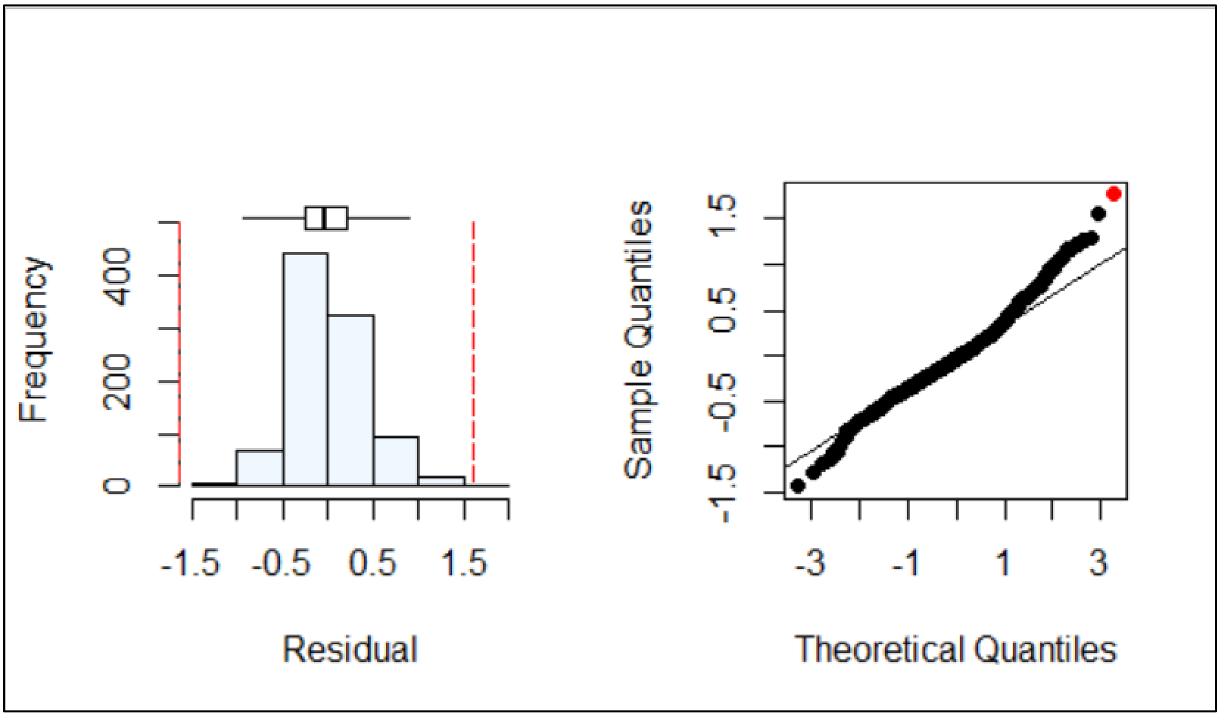


Figure 3: Histogram and quantile plot of residuals from exploratory model

Table 3: REML estimates of parameters and AIC for the exploratory model, null model and the hypothesis tests.

Test	Predictors	Partial Sill	Range	Nugget	AIC	
					Max.likelihood	Non-spatial
	Exploratory (all predictors)	0.1367	21.08	0.218	1570	1664
0	mean	0.259	68.28	0.224	1618	1946
1	Rainfall	0.217	51.87	0.222	1611	1865
2	Rainfall + Soilclass	0.197	52.71	0.221	1622	1817
3	Rainfall +Slope	0.214	51.98	0.222	1610	1856
4	Rainfall + Elevation	0.180	39.70	0.220	1598	1774
5	Rainfall + Elevation + Valley depth	0.173	33.48	0.218	1598	1770
6	Rainfall+ Elevation + Landcover	0.174	40.09	0.220	1603	1769
7	Rainfall + Elevation + NDVI	0.181	39.83	0.220	1600	1776
8	Rainfall + Elevation + LS	0.174	38.84	0.221	1596	1760

9	Rainfall + Elevation + LS + CNBL	0.171	37.40	0.221	1597	1783
10	Rainfall + Elevation + LS + RSP	0.171	38.13	0.221	1598	1751
11	Rainfall + Elevation + LS + Aspect	0.173	39.18	0.221	1598	1760

Table 4 shows the log likelihood ratio and p-values of each test at respective degree of freedom df and chi-square distribution values. The likelihood ratios of tests 1,4 and 8 were greater than the chi-square distribution value values. Therefore, the null hypothesis for these cases were rejected, and the predictors retained during the sequential testing. The rest of the tests had log likelihood ratio less than their respective chi-square distribution values. Hence the null hypothesis was accepted for these predictors and they were dropped.

Table 4: likelihood ratio and p-values of each hypothesis test at respective degree of freedom(df) and chi-square distribution values.

Test		**df**	**Chi-square**	**Likelihood ratio**	**p-value**
1	Rainfall	1	3.841	9.533	0.002
2	Rainfall + Soilclass	17	27.587	22.506	0.166
3	Rainfall +Slope	1	3.841	2.429	0.119
4	Rainfall + Elevation	1	3.841	14.416	0.000
5	Rainfall + Elevation + Valley depth	1	3.841	1.817	0.177
6	Rainfall+ Elevation + Landcover	8	15.507	10.836	0.211
7	Rainfall + Elevation + NDVI	1	3.841	0.598	0.439
8	Rainfall + Elevation + LS	1	3.841	3.946	0.047
9	Rainfall + Elevation + LS + CNBL	1	3.841	1.047	0.306
10	Rainfall + Elevation + LS + RSP	1	3.841	0.403	0.525
11	Rainfall + Elevation + LS + Aspect	1	3.841	0.213	0.644

Df= Degree of freedom which is the difference between the total degree of freedom for the target model and that of the model being compared to, LS= LS-Factor, CNBL= Channel Network Base Level, RSP= relative slope position.

Figure 4.a shows the alpha wealth after each test and it can be observed that the quantity of the wealth is increased when the null hypothesis is rejected and depleted when the null hypothesis retained, and it goes to zero at the end of the sequence. Figure 4.b shows the p-values (open symbols) for the successive tests of additional predictors, as in Table 4, and the threshold (solid symbols) against which each successive p-value is tested to achieve FDR. On this basis rainfall and elevation were selected as predictors.

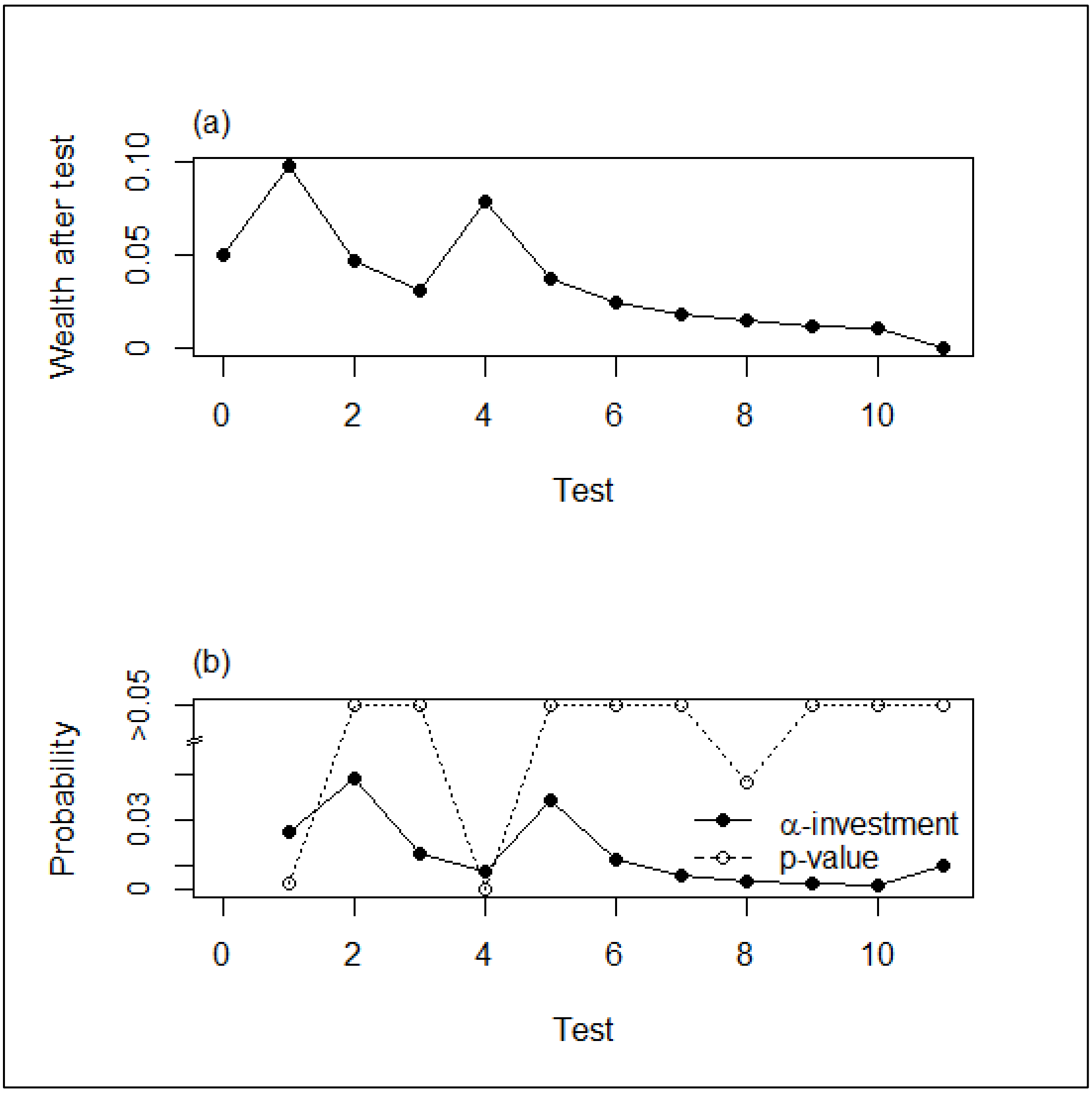


Figure 4: a. alpha wealth after each test. b. probability of alpha investment and p-values

4.2 Spatial prediction of soil pH

The estimated covariance parameters for the linear mixed models to be used for spatial prediction of soil pH by the E-BLUP with elevation and rainfall as fixed effects for prediction (Method A) and a constant mean as the only fixed effect (Method B, equivalent to ordinary kriging) are shown in Table 5. The nugget, partial sill and range for the model with rainfall and elevation as fixed effects are 0.220, 0.195 and 33.95 respectively. These values are smaller than the corresponding parameters for the model with a constant mean as the only fixed effect (0.224, 0.269 and 72.83). AIC values for both models are less than those of respective non-spatial AIC. On this basis we may conclude that there is evidence for spatial dependence in the random component of the LMM, and so potentially benefits in computing the E-BLUP for spatial prediction at unsampled sites.

Table 5: Covariance parameters for A= REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), B=REML-EBLUP with the only fixed effect a constant mean (ordinary kriging).

Method	**Partial Sill**	**Range**	**Nugget**	**AIC**	
				Max.likelihood	**Non-spatial**
A	0.195	33.95	0.220	1184	1310
B	0.269	72.83	0.224	1614	1945

Table 6 shows the number of trees, number of predictors, number of variables considered at each split, target node size and out-of-bag cross validation of the two random forest methods. The out-of-bag MSE and R-squared show that there is a slight reduction in performance of the random forest model with rainfall and elevation

Table 6: ntree = number of trees in the forest; mtry = number of variables considered at each split

Method	**ntree**	**predictors**	**mtry**	**Target node size**	**Out-of-Bag MSE**	**Out-of-Bag R-squared**

C	200	11	3	**5**	0.31	0.30
D	200	2	1	5	0.32	0.27

Table 7 shows the permutation variable importance for each predictor when a random forest model is fit with all predictors alone and when we include null predictors (sim1 to sim6) which were generated by simulation to examine how random forest variable importance performs with predictors that have no relation to the data. For the random forest model, the most important variable is elevation with importance value of 0.166, followed by Channel Network Base Level with value of 0.155. some variable importance values for some predictors are almost equal or even less, but their p-values are much smaller. Null variables sim1 and sim6 despite have low variable importance values, but very small p-values (less than 0.01). The inclusion of these null variables has a substantial effect on the p-values of some predictors such as soilclass, slope, landcover.

Table 7: Permutation variable importance and p-values when a random forest model is fit with all predictors alone and when we include null predictors (sim1 to sim6).

Predictor	No null predictors		Null predictors included	
	Importance	p-value	Importance	p-value
rain	0.0889	0.0099	0.0912	0.0099
soilclass	0.0231	0.0198	0.0153	0.0099
slope	0.0318	0.8218	0.0268	0.2079
elevation	0.1657	0.0099	0.1718	0.0099
valley	0.0761	0.0099	0.0544	0.0099
landcover	0.0063	0.4752	0.0074	0.0792
NDVI	0.0499	0.0099	0.0470	0.0099
ls	0.0467	0.3267	0.0286	0.1287
cnbl	0.1554	0.0099	0.1477	0.0099
rsp	0.0554	0.0198	0.0330	0.0495
aspect	0.0148	0.1188	0.0066	0.3663
Sim1			0.0279	0.0099
Sim2			0.0232	0.0198
Sim3			0.0187	0.0594
Sim4			0.0204	0.0198
Sim5			0.0160	0.1782
Sim6			0.0274	0.0099

Table 8 shows the estimated parameters of the random forest residuals for the exponential, spherical and pure nugget correlation models. the non-spatial model was preferred because the AIC values for the spatial component was higher than that of the non-spatial component in both random forest predictions. Indeed, the fitted correlated variance for the spatial covariance function was zero. On this basis there is no scope to improve the RF predictions by a kriging step.

Table 8: Estimated parameters of the exponential, spherical and pure nugget correlation functions for the residuals of the two random forest predictions.

Method	Parameter	Exponential	Spherical	Pure.nugget
RF (dem + rain)	Partial Sil	0	0	0.188
	range	0	0	50
	Nugget	0.188	0.188	0
	AIC$_{\text{max.lelihood}}$	1123	1123	1123
	AIC$_{\text{non-spatial}}$	1119	1119	1119
RF (all predictors)	Partial Sil	0	0	0.157
	range	0	0	50
	Nugget	0.157	0.157	0
	AIC$_{\text{max.likelihood}}$	951.4	951.4	951.5
	AIC$_{\text{non-spatial}}$	947.4	947.4	947.4

Figure 5 shows the predicted spatial variability of soil pH. The spatial pattern is similar for all the models with low pH values (less than 5.5) in the Western and Northern parts and higher values (above 6) in the Southern and Eastern parts.

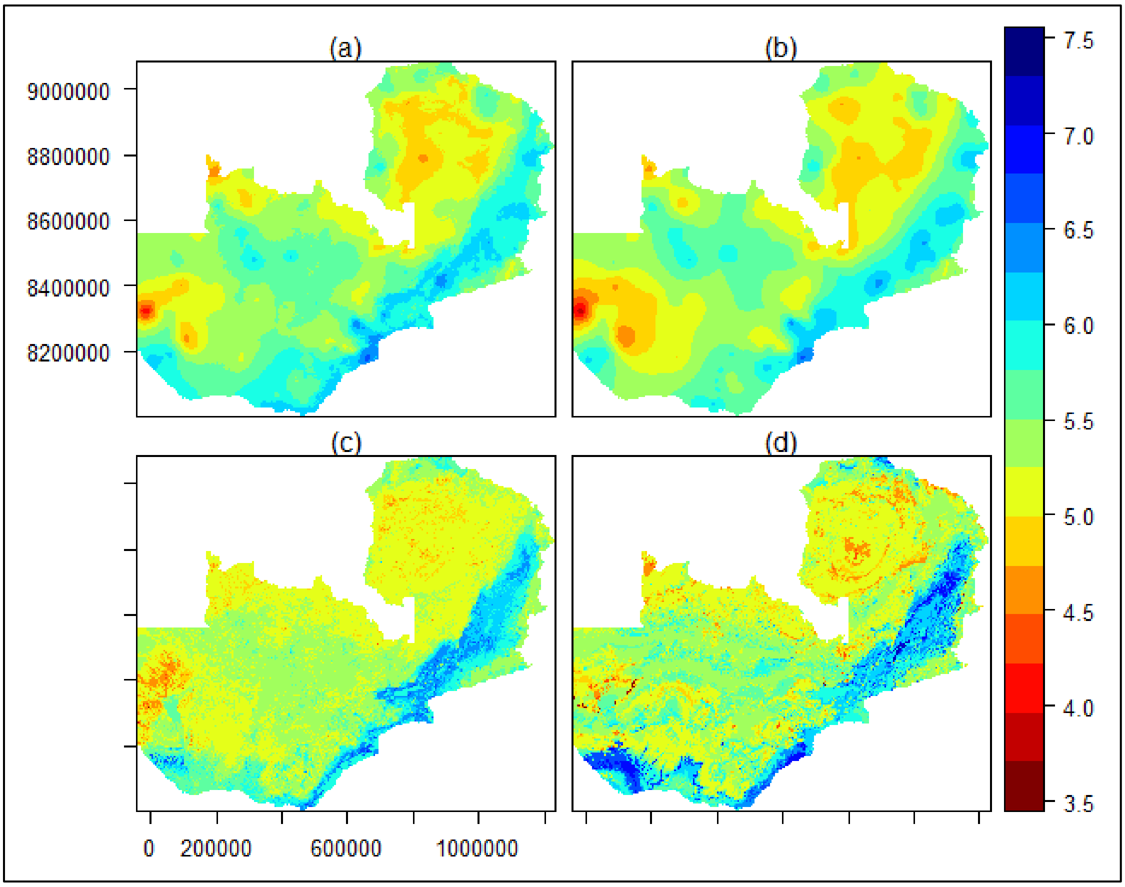


Figure 5: Prediction maps of soil pH for (a) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), (b)REML-EBLUP with the only fixed effect a constant mean (ordinary kriging (c) random forest with all predictors (d). random forest with elevation and rainfall as predictors selected through alpha-investment.

4.3 Map Validation

Table 9 shows the summary validation statistics for (A) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), (B)REML-EBLUP with the only fixed effect a constant mean (ordinary kriging), (C) random forest with all predictors (D). random forest with elevation and rainfall as predictors selected through alpha-investment. The mean and median error values were smallest for the REML-EBLUP (ordinary kriging) method while that of the REML-EBLUP (kriging with external drift) was larger than that of the two random forest methods. The MSE and RMSE for the two REML-

EBLUP methods were smaller than those of the two random forest methods with REML-EBLUP (ordinary kriging) having the smallest values. There was spatial dependency in the prediction error in all the cases with the two REML-EBLUP cases having the smaller partial sill values of 0.15 compared to 0.30 and 0.22 values for Random forest (elevation and rainfall predictors) and random forest (all predictors) respectively. The ESE values for all the cases were larger than the MSE values because the bias (ME) for the models is greater than zero. REML-EBLUP (ordinary kriging) had the smallest ESE value.

Table 9: Summary Validation statistics for(A) REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), (B)REML-EBLUP with the only fixed effect a constant mean (ordinary kriging), (C) random forest with all predictors (D). random forest with elevation and rainfall as predictors selected through alpha-investment.

Variable		**A**	**B**	**C**	**D**
Prediction error	Mean	0.168	0.094	0.116	0.128
	Median	0.212	0.148	0.200	0.200
MSE		0.417	0.388	0.463	0.551
Corr.Model		Exponential	Exponential	Exponential	Exponential
Partial Sil		0.154	0.145	0.218	0.299
Range		48.380	44.670	40.200	36.250
Nugget		0.240	0.240	0.237	0.238
ESE		0.422	0.393	0.468	0.553

5. DISCUSSION

The mapped soil pH by all approaches is shown in Figure 5. The optimum pH ($CaCl_2$) for plant growth is between 5.2 – 7.5. bellow the pH of 5.2, the levels of Aluminum, Manganese and Copper are toxic for plant growth, Phosphorous and Magnesium are not available to plant. Above pH of 7.5, the interactions between Calcium, Magnesium and Potassium have a negative impact on root absorption. Copper, Iron, Manganese, Zinc, Boron and Phosphorous are deficient (Lake, 2000). The maps in Figure 5 show pH values less than 5.2 in the western and northern parts of the country meaning we expect these areas to have challenges of Aluminum, manganese and copper toxicity as well as Phosphorous and Magnesium deficiencies. In the Southern parts of the country the pH values, according to all the maps in Figure 5, range from 5.2 to 7.5 which are optimal for plant growth. There are few areas in the southern part of Zambia with pH above 7.5. Similar spatial variations were observed by Chapoto et al., 2016. The southern parts where the pH is high is a valley area, the northern parts receives high rainfall and the western part despite receiving the same amount of rainfall as the eastern part, the areas is characterized by Kalahari sand. Our results show a similar spatial pattern for soil pH as that presented in the SoilGrids map (www.soilgrids.org) of Hengl et al., (2017). The main difference is that our map shows low pH values in the west of the country, whereas the SoilGrids map shows larger values there. Our results are more plausible pedologically given the parent material, and it has been long-established that the soils formed over the Kalahari sands of western Zambia are weakly to extremely acidic (Brammer, 1976). A more thorough assessment of the SoilGrids predictions using the RALS data would be of interest.

Predictions by the E-BLUP from the LMM with the only fixed effect a constant mean (equivalent to ordinary kriging) were better than other predictions in the sense that the mean and median errors were closest to zero and the mean square error and expected square error were the smallest. This is unexpected, given the evidence provided in the model-fitting stage for a significant relationship between soil pH and the selected covariates. This might be expected to result in better predictions from the LMM which includes these covariates as fixed effects. However, one may note (Table 5) that the correlated random variance in the LMM with rainfall and elevation as fixed effects is only about 25% smaller than the corresponding variance in the LMM with a constant mean the only fixed effect. The fact that a covariate is significantly related to a soil property does not necessarily mean that it will allow improved prediction of that property relative to a model without that covariate. That is because the corresponding fixed effect coefficient must be estimated, and this estimation is a source of error in the prediction. Furthermore, Zimmerman et al., (1999) found that ordinary kriging performed better than universal kriging (UK, equivalent to the E-BLUP with some covariates) with a spatially clustered data set, while UK performed better when the data were not clustered. This may be because, in a strongly clustered data set, the effective degrees of freedom with which the fixed effects coefficients are estimated may be relatively small.

The use of random forests to include the environmental covariates in spatial prediction was less successful than the LMM and E-BLUP, with larger values of ESE. This could be due to over-fitting. It is notable that the residuals from the fitted RF at the calibration data points showed no spatial dependence, while the RF prediction errors at the validation points (Table 9) do show spatial dependence. This could arise because the RF algorithm, given its flexibility and ability to fit non-linear relationships, generates a model which closely fits the variations within the training data set, but which is not representative of the relationship between the predictor variables and target variable in the underlying population. This would lead both to marked bias in models of the random variation based on the residuals, as can also occur with ordinary least squares (Lark et al., 2006) and also in poor performance of the RF on a separate validation data set. These data may also provide a problem for the RF methodology because of the strong spatial clustering. If some data from a cluster are used in the development of trees while others are in the OOB subset then the assessment of the model and the value of the predictors may be over-optimistic. A predictor variable overfitted to a clustered data set might well fail to predict effectively at independent validation points. This emphasizes the importance of a genuinely independent validation of spatial predictions (Brus et al., 2011).

Spatial clustering of the observations may also be a contributing factor to the small p-values attributed to the entirely random, although spatially autocorrelated, null predictors which we evaluated. This gives reason for caution when interpreting RF output. It is consistent with the findings of Wadoux et al., (2020) that the RF algorithm may select as predictor variables covariates which are not related to the target properties of interest by any direct or indirect causal relations. A spatially dependent predictor variable of this nature may indeed support spatial prediction of a variable to which it has no underlying relationship, but if this is the case then one might prefer to use a properly-designed set of orthogonal polynomial basis functions for the model rather than arbitrary variables. Furthermore, with strongly spatially clustered data, it is even more likely that a uncausally-related predictor will result in poor predictions at independent validation sites.

Many digital soil mapping studies use legacy data sets, rather than new samples collected for the purpose. As legacy data sets may originate in local surveys, or from networks of experimental stations, they may show marked spatial clustering, as do the RALS data because of their two-stage cluster sampling design. We note that such clustering may cause difficulties for the RF algorithm but that it is also important to account for it when dividing data into prediction and validation subsets. There is a risk of bias in the validation of a map if validation and training data are drawn from common clusters. The estimation of validation statistics from a validation set which is strongly clustered may also result in bias, which is why we have used a model-based approach to compute these statistics in this study. The expected squared error, computed from the model, in this case is not very different from the mean squared error computed directly although in each case it exceed the mean square error as expected (Table 9). This could be because the clusters are reasonably balanced (similar numbers of observations in each), and are themselves selected independently and at random. The model-based method to quantify prediction uncertainty is, nonetheless, a more general approach for use with validation data from locations not selected by probability sampling.

6. CONCLUSION

Spatial variability of soil pH was mapped using REML-EBLUP with elevation and rainfall as fixed effects selected through alpha-investment (kriging with external drift), REML-EBLUP with the only fixed effect a constant mean (ordinary kriging), random forest with all predictors and random forest with elevation and rainfall as predictors selected through alpha-investment models. The soil pH maps from these models showed similar patterns with pH values less than 5.2 in the Western and Northern parts of the country. In the Southern parts of the country the pH values range from 5.2 to 7.5 which are optimal levels of soil pH for plant growth.

The ME, MSE and ESE (computed as the sum of the squared mean error and the nugget and spatially correlated variances from the LMM) were used to validate the performance of the models for spatial prediction of soil pH. The values of the ME, MSE and ESE from the validation statistics showed that REML-EBLUP with the only fixed effect a constant mean (ordinary kriging) performed better than the other methods.

Random forests had the largest values of MSE and ESE. This may result from over-fitting, and from the strongly spatially clustered distribution of the observations in the legacy data set which could affect the internal cross-validation in the RF algorithm.

We also noticed that the algorithm appeared susceptible to wholly random "null predictors" which we had simulated. Other studies, notably by Wadoux et al. (2020) have shown this, but we believe this to be the first example where mutually independent random spatially autocorrelated candidate predictor variables have been selected alongside pedologically plausible ones. The selection of such null predictors should give pause as it suggests that the random forest algorithm may be prone to overfitting. We suggest that this problem warrants further study as pedometricians should always aim to generate insight from their analyses, as well as predictions.

We note that legacy data, often used in digital soil mapping, may be strongly clustered like ours. We emphasize again the importance of splitting data into prediction and validation subsets at cluster level (i.e. allocating all data in any one cluster to the prediction or to the validation set). In this case there was not a large difference between the ESE (model-based estimate of the expected squared error) and the MSE (average of the squared errors), but this would not be true in general, and the use of a model-based approach to the analysis of validation errors at locations not selected independently and at random is most appropriate.

Finally, we note that we found no evidence for spatial correlation in the residuals from the fit of the random forest to the prediction data set, although there was correlation in the prediction errors by this method at the validation sites. This is an important reminder that such residuals given us little if any insight into the actual behavior of the prediction error, and are good reasons to avoid using kriging methods in combination with modelling methods which, unlike REML and EBLUP, do not have a built-in methodology to estimate parameters of the error without bias.

7. ACKNOWLEDGEMENT

This study was made possible with support from the Zambian Ministry of Higher Education through their 2016/2017 science and technology female postgraduate scholarships; and from the Commonwealth Scholarships Commission through their split-site scholarship funded by the UK government. We also thank Indaba Agricultural Policy Research Institute (IAPRI) for allowing us to use their RALS 2012 survey data.

8. REFERENCES

Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure, Bioinform. 26,1340-1347, https://doi.org/:10.1093/bioinformatics/btq134

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal., 52(4), .2249-2260, https://doi.org/10.1016/j.csda.2007.08.015

Behrens, T., Scholten, T., 2007. A Comparison of Data-Mining Techniques in Predictive Soil Mapping, in: Lagacherie, P., McBratney, A, B., Voltz, M. (Eds.), Developments in Soil Sci. Elsevier B.V., 353.

Brady, N.C. and Weil, R.R., 2014. The nature and properties of soil. (14th New Int. Ed.).

Brammer, H. 1976. Soils of Zambia. Department of Agriculture, Land Use Branch. Lusaka.

Breiman, L., 1996. Bagging Predictors. Mach. learn., 26(2), 123-140.

Breiman, L., 2001. Random forests. Mach. learn., 45(1), 5-32.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62(3), pp.394-407.

Burke, W.J., Frossard, E., Kabwe, S., Jayne, T.S., 2019. Understanding fertilizer adoption and effectiveness on maize in Zambia. Food policy, 86, 101721, https://doi.org/10.1016/j.foodpol.2019.05.004

Central Statistical Office, Ministry of Agriculture and Livestock, and Indaba Agricultural Policy Research Institute (CSO/MAL/IAPRI)., 2015. Rural Agricultural Livelihoods Survey. Lusaka, Zambia: CSO/MAL/IAPRI.

Central Statistical Office (CSO)., 2012. Rural Agricultural Livelihoods Survey; Instruction Manual for Listing, Sample Selection, and Largest Maize Field Data Collection**.**

Chai, X., Shen, C., Yuan, X., Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma, 148(2), 159-166, https://doi.org/10.1016/j.geoderma.2008.09.018

Chapoto, A., Chabala, L. M., Lungu, O. N., 2016. "A Long History of Low Productivity in Zambia: Is it Time to Do Away with Blanket Recommendations?" Zambia Social Sci. J. 6: (2), Article 6.

Cummings, M.P., Myers, D.S., Mangelson, M., 2004. Applying permutation tests to tree-based statistical models: extending the R package rpart. In Tech Rep CS-TR-4581, UMIACS-TR-2004-24, Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland.

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

ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. 2017. http://maps.elie.ucl.ac.be/CCI/viewer/download.php (Accessed on 20th July 2019).

Fick, S.E., Hijmans R.J., 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land area. Intl. J. Climatol. 37 (12): 4302-4315. https://www.worldclim.org

Foster, D.P., Stine, R.A., 2008. α-investing: a procedure for sequential control of expected false discoveries. J. R. Stat. Soc.: Series B (Stat. Methodol.), 70(2), 429-444.

Gashu, D., Lark, R.M., Milne, A.E., Amede, T., Bailey, E.H., Chagumaira, C., Dunham, S.J., Gameda, S., Kumssa, D.B., Mossa, A.W., Walsh, M.G., 2020. Spatial prediction of the concentration of selenium (Se) in grain across part of Amhara Region, Ethiopia. Sci. Total Environ. p.139231.

Government of the Republic of Zambia (GRZ). 1991. Exploratory Soil Map of Zambia (1: 1.000, 000).

Government of the Republic of Zambia (GRZ) and UNDP., 2009. Adaptation to the effects of drought and climate change in Agro-ecological Regions I and II in Zambia. Project document. Ministry of Agriculture and Cooperatives. Accepted by: Ministry of finance and National Planning, and UNDP.

Hengl, T., 2003. Pedometric mapping: bridging the gaps between the conventional and pedometric approaches. Wageningen University.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS One 10, 1–26, https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, *6*, p.e5518, https://doi.org/10.7717/peerj.5518

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O., Schad, P., 2006. Guidelines for soil description. FAO.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning 112, 3-7. New York: springer, https://doi.org/10.1007/978-1-4614-7138-7.

Jones Jr, J. Benton., 2012. Plant nutrition and soil fertility manual. CRC press.

Kempen, B., Heuvelink, B, G., Brus, J, D., Stoorvogel, J, J., 2010. Pedometric mapping of soil organic matter using a soil. Eur. J. Soil Sci. 61, 333–347, https://doi.org/10.1111/j.1365-2389.2010.01232.x

Kitanidis, P.K. 1987. Parametric estimation of covariances of regionalized variables. Water Resour. Bull., 23, 557–567.

Kienast-Brown, Suzann Libohova, Z., Janis, B., 2010. Digital Soil Mapping, in: Soil Survey Manual. USDA-NRCS, 429–436, https://doi.org/10.1007/978-90-481-8863-5

Lake, B., 2000. Understanding soil pH. Acid Soil Action. Leaflet, (2).

Lark, R.M., 2017. Controlling the marginal false discovery rate in inferences from a soil dataset with α-investment. Eur. J. Soil Sci. 68(2),221-234.

Lark, R. M., Ander, E. L., Broadley, M. R. 2019. Combining two national-scale datasets to map soil properties, the case of available magnesium in England and Wales. Eur. J. Soil Sci. 70(2), 361-377.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. Eur. J. Soil Sci. *55*(4),799-813, https://doi.org/10.1111/j.1365-2389.2004.00637.x

Lark, R.M., Cullis, B.R., Welham, S.J. 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57(6): 787-799, https://doi.org/10.1111/j.1365-2389.2005.00768.x

Lark, R.M., Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surface Processes and Landforms: The J. Br. Geomorphol. Res. Group, 31(7), 862-874, https://doi.org/10.1002/esp.1296

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), 18-22.

Li, J., Heap, A.D., Potter, A., Huang, Z., Daniell, J.J., 2011. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. Cont. Shelf Res. 31(13), 1365-1376.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey q. Geoderma 97, 293–327.

McCauley, A., Jones, C., Jacobsen, J., 2009. Soil pH and organic matter. Nutr. Manag. Modul. 8(2), 1-12.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma, 140(4), 324-336, https://doi.org/10.1016/j.geoderma.2007.04.028

Moonjun, R., Farshad, A., Shrestha, D, P., Vaiphasa, C., 2010. Artificial Neural Network and Decision Tree in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand, in: Boettinger, J, L., Kienast-Brown, S., Howell, D, W., Moore, A, C., Hartemink, A, E. (Eds.), Digital Soil Mapping (Bridging Research, Environmental Application, and Operation). Springer Science+Business Media B.V.

Omutu, T. C., 2020. soilassessment: Assessment Models for Agriculture Soil Conditions and Crop Suitability, https://CRAN.R-project.org/package=soilassessment

Patterson, H. D., Thompson, R., 1971. Recovery of inter block information when block sizes are unequal. Biom. 58, 545-554.

Rawlins, B.G., Lark, R.M., O'donnell, K.E., Tye, A.M., Lister, T.R., 2005. The assessment of point and diffuse metal pollution of soils from an urban geochemical survey of Sheffield, England. Soil Use Manag. 21(4), 353-362, https://doi.org/10.1079/SUM2005335

Ribeiro, P. J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R-News 1, 15–18.

Schlather, M., Malinowski, A., Menck, P.J., Oesting, M., Strokorb, K., 2015. Analysis, simulation and prediction of multivariate random fields with package random fields. J. Stat. Softw. 63(8), 1-25, https://doi.org/10.18637/jss.v063.i08

Stein, M.L. 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinform. 8(1), 25, https://doi.org/10.1186/1471-2105-8-25

Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12(10), 1687, https://doi.org/10.3390/rs12101687

Swallow, W. H., Monahan, J. F., 1984. Monte-Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components. Technometr. 26, 47-57

United States Geological Survey (USGS)., 2019 NASA Shuttle Radar Topography Mission (SRTM3) data available on the World Wide Web https://earthexplorer.usgs.gov (accessed 10 July, 2019)

Veldkamp, W.J., Muchinda, M. and Delmotte, A.P., 1984. Agro-climatic zones in Zambia. Soil Survey Bulletin (Zambia).

Verbeke, G., Molenberghs, G., 2000. Linear mixed models for longitudinal data. Springer-Verlag, New York.

Viscarra Rossel, R.A., Webster, R., Kidd, D., 2014. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf. Process. Landf., 39(6), 735-748.

Wadoux, A.M.C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020. A note on knowledge discovery and machine learning in digital soil mapping. Eur. J. Soil Sci. 71(2), 133-136.

Webster, R. and Oliver, M.A., 2007. Geostatistics for environmental scientists. John Wiley & Sons.

Wright, M. N., Ziegler, A., 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1-17, https://doi.org/10.18637/jss.v077.i01

Zimmerman, D.L., Zimmerman, M.B., 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometr., 33(1), 77-91, https://doi.org/10.1080/00401706.1991.10484771

Zimmerman, D., Pavlik, C., Ruggles, A., Armstrong, P, M., 1999. An Experimental Comparison of Ordinary and Universal Kriging and Inverse Distance Weighting. Math. Geol. 31, 375–390, https://doi.org/10.1023/A:100758650743