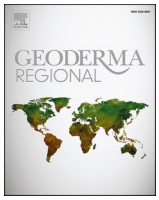
Geoderma Regional 35 (2023) e00731

Contents lists available at ScienceDirect

Geoderma Regional

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



Predictive geochemical mapping using machine learning in western Kenya

Olivier S. Humphrey [a,*], Mark Cave [a], Elliott M. Hamilton [a], Odipo Osano [b], Diana Menya [c], Michael J. Watts [a]

[a] Inorganic Geochemistry, Centre for Environmental Geochemistry, British Geological Survey, Nottingham, UK
[b] School of Environmental Sciences, University of Eldoret, Eldoret, Kenya
[c] School of Public Health, Moi University, Eldoret, Kenya

ARTICLE INFO

Keywords:
Random Forest
Machine learning
Soil
Geochemistry
Uncertainty
Kenya

ABSTRACT

Digital soil mapping techniques represent a cost-effective method for obtaining detailed information regarding the spatial distribution of chemical elements in soils. Machine learning (ML) algorithms using random forest (RF) models have been developed for classification, pattern recognition and regression tasks, they are capable of modelling non-linear relationships using a range of datasets, identifying hierarchical relationships, and determining the importance of predictor variables. In this study, we describe a framework for spatial prediction based on RF modelling where inverse distance weighted (IDW) predictors are used in conjunction with ancillary environmental covariates. The model was applied to predict the total concentration (mg kg^{-1}) and assess the prediction uncertainty of 56 elements, soil pH and organic matter content using 466 soil samples in western Kenya; the results of iodine (I), selenium (Se), zinc (Zn) and soil pH are highlighted in this work. These elements were selected due to contrasting biogeochemical cycles and widespread dietary deficiencies in sub-Saharan Africa, whilst soil pH is an important parameter controlling soil chemical reactions. Algorithm performance was evaluated determining the relative importance of each predictor variable and the model's response using partial dependence profiles. The accuracy and precision of each RF model were assessed by evaluating out-of-bag predicted values. The models R^2 values range from 0.31 to 0.64 whilst CCC values range from 0.51 to 0.77. The IDW predictor variables had the greatest impact on assessing the distribution of soil properties in the study area, however, the inclusion of ancillary environmental data improved model performance for all soil properties. The results presented in this paper highlight the benefits of ML algorithms which can incorporate multiple layers of data for spatial prediction, uncertainty assessment and attributing variable importance. Additional research is now required to ensure health practitioners and the agri-community utilise the geochemical maps presented here for assessing the relationship between environmental geochemistry, endemic diseases and preventable micronutrient deficiency.

1. Introduction

Digital soil mapping (DSM) employs a generic framework to predict a target variable or class at an unobserved location based on the quantitative relationship between georeferenced observations and one or more environmental covariate which is likely to impact the variable or class of interest within a defined area (Asgari et al., 2020; Lagacherie et al., 2006; McBratney et al., 2003; Sylvain et al., 2021; Wadoux et al., 2019; Zeraatpisheh et al., 2019). The spatial distribution of chemical elements in soils, originating from geogenic and anthropogenic sources, can provide critical information for assessing mineral exploration, environmental monitoring, and assessing nutrient dynamics (Hengl et al., 2015; Johnson and Ander, 2008; Sylvain et al., 2021; Wadoux et al., 2019). The need for comprehensive, accurate and up-to-date soil information maps is an essential component for the formulation of agricultural policies, soil management strategies, and monitoring environmental impact arising from changing land use (Chagas et al., 2016; Hengl et al., 2015; Towett et al., 2015).

Geochemical maps can be used to investigate the relationship between environmental geochemistry and endemic diseases. Esophageal cancer (EC) has a unique spatial distribution, particularly for the histological subtype of esophageal squamous cell carcinoma (ESCC) which predominates across central Asia and along the eastern Africa corridor extending from Ethiopia to South Africa (Schaafsma et al., 2015). Whilst

* Corresponding author.
E-mail address: olih@bgs.ac.uk (O.S. Humphrey).

https://doi.org/10.1016/j.geodrs.2023.e00731
Received 23 September 2022; Received in revised form 16 October 2023; Accepted 22 October 2023
Available online 25 October 2023

co-presenting factors may explain the occurrence of this disease in young African patients (<30 years), the risk of ESCC was linked to suboptimal nutrient intake (Liu et al., 2013); however, evidence for the role of specific nutrients is not conclusive. Nevertheless, there is evidence of an inverse association between selenium (Se) and zinc (Zn) in ESCC risk studies (Hashemian et al., 2014; Steevens et al., 2010). Micronutrient deficiencies are common in Africa, particularly for Zn, Se, and iodine (I), the prevalence of which has been shown to have large spatial variation (Joy et al., 2014). The variation of nutrient availability to crops is influenced by a large number of factors including soil properties such as pH. In Malawi, the reduced concentration of Se in crops grown on non-calcareous soils (pH < 6.5) compared to calcareous soils (pH > 6.5) was observed due to the limited phyto-availability of Se in acidic soils (Joy et al., 2015). Soil properties also affect I fixation, mobility, and speciation (Humphrey et al., 2020). Previous studies in western Kenya have highlighted that up to 12% of the population were potentially I deficient, while over 44% were considered to have an excess I intake (Watts et al., 2020). Additional research investigating the source apportionment of micronutrients in the diets of western Kenyans used urinary biomonitoring to provide a snapshot of population micronutrient status (Watts et al., 2019). The overall calculated deficiency rates were close to or exceeded 90% for Se and Zn. Due to the strong reliance on locally grown food products, which are influenced by local soil properties, any elemental excess or deficiencies would predominantly or exclusively originate from local sources (Schaafsma et al., 2015). Recently, Gashu et al. (2021) identified geographic hotspots of human micronutrient deficiency associated with the geospatial variation in the composition of micronutrients in crops in Ethiopia and Malawi. They highlighted that the location of rural households consuming locally sourced food is the largest influencing factor in determining the dietary intake of micronutrients from cereals. This makes soil geochemistry a crucial factor in understanding the spatial incidence of endemic diseases and the prevalence ESCC in Eastern Africa.

Conventionally, spatial predictions of soil properties have been performed by kriging and its many variants (Goovaerts, 1999; Minasny and McBratney, 2007; Oliver and Webster, 1990; Zhu and Lin, 2010). Kriging is a linear unbiased predictor and one of the most applied spatial interpolation techniques for estimating soil properties as it also provides a measure of the probable error associated with the estimates (Webster and Oliver, 2007). Despite its advantages, kriging has several disadvantages which have only been partially addressed. Kriging makes several assumptions that residuals are normally distributed, stationary (with constant mean and unit variance) and isotropic and incorporating supplementary cross-correlated covariates introduces further challenges (Wadoux et al., 2020). In heterogeneous regions, the model fails to encapsulate gradual and rapid changes in soil variation, and dependent upon the sample size and prediction area, kriging can present significant computationally challenges (Wadoux et al., 2020).

In recent years (supervised) machine learning (ML), referring to a large class of non-linear data-driven algorithms, have been applied to data mining, pattern recognition, regression and classification tasks (Hengl et al., 2018; Wadoux et al., 2019). First described by Breiman (2001) random forest (RF) is a hierarchical nonparametric ML algorithm that consists of a large number of individual tree models trained from bootstrap samples and has proven to be efficient for producing spatial predictions. The decision trees classify data by inferring the relationships between a dependent variable and a set of predictors before the RF model aggregates the results of all individual trees to make a single prediction (Naimi et al., 2022; Pouladi et al., 2019). This approach has several advantages including its ability to model non-linear relationships, using numerical, ordinal, binary, and categorical datasets; reduce potential overfitting and bias, identifying complex hierarchical relationships between predictors and response variables; and relationships between predictors providing the relative importance of a predictor variable based on the regression prediction error of out-of-bag (OOB) predictions (Heung et al., 2014; Pouladi et al., 2019). Despite the increased use of the RF frameworks for spatial interpolation most do not consider that the observations are geo-referenced and may be spatially correlated (Sekulić et al., 2020). Hengl et al. (2018) introduced a RF for spatial predictions framework (RFsp) where Euclidean buffer distances from observation points are used as explanatory variables, thus incorporating geographical proximity effects into the prediction process to mimic spatial correlation used in kriging, it was noted that adding these covariates improved prediction. Cave (2017) and Sekulić et al. (2020) applied similar approaches using inverse distance weighted (IDW) and neighbouring observations as predictor variables, respectively, in RF models with environmental covariates to improve the models prediction performance. Inverse distance weighted (IDW) interpolation assumes that samples that are close to one another are more alike than those that are further apart (Fortin and Dale, 2006). Cave (2017) highlighted that the use of IDW covariates displays more structure in the data as it better considers localised variability in comparison to ordinary kriging, which uses a single variogram to estimate point weighting for the prediction area. Similarly to kriging, the RF methods can also produce prediction variance estimates derived from the bootstrap sampling of the data points (inherent in the RF algorithm). This approach has subsequently been used to predict the spatial distribution of persistent organic pollutants in London (Vane et al., 2021), and produce spatial prediction models of the total and bioaccessible fractions of arsenic and lead in an urban environment (Wragg and Cave, 2021).

The aim of this study was to use random forest modelling, utilising both IDW and ancillary environmental covariates, to spatially predict soil properties and assess their prediction uncertainty in western Kenya. The objectives of this study were to (1) develop optimised RF models using a feature ranking algorithm to identify statistically significant variables; (2) use a non-parametric post-processing tool to explain the importance of significant attributes; and (3) compare the performance of different soil mapping techniques. The model was applied to predict the total concentration (mg kg^{-1}) and assess the prediction uncertainty of 56 elements, soil pH and organic matter content in 466 soil samples from western Kenya, the open-access database is available in Watts et al. (2021a). The results of iodine (I), selenium (Se), zinc (Zn) and soil pH are highlighted in this work. These elements were selected due to their contrasting biogeochemical cycles and reported widespread deficiencies, while soil pH was assessed due to its profound impact on soil chemical reactions influencing elemental fixation, mobility, and speciation.

2. Materials and methods

2.1. Study area

The soil geochemical data used in this study were derived from samples collected between October 2016 and November 2019 as part of a wider project described in Watts et al. (2019) which collected soil, crop, drinking water and urine samples from household 'shambas' (produce plots) in rural locations to estimate micronutrient intakes and subsequent risk of deficiencies. Topsoil (0–15 cm) samples were collected from 446 sampling sites in 15 western Kenya counties, including Bomet, Bungoma, Busia, Elgeyo Marakwet, Homa Bay, Kakamega, Kericho, Kisii, Kisumu, Nandi, Nyamira, Siaya, Trans Nzoia, Uasin Gishu, and Vihaga (Fig. 1).

2.2. Ethical approval

For the wider study, which included human biomonitoring (Watts et al., 2021b), ethical approval was obtained from Moi University Institutional Research Ethics Committee (IREC 000921). Permission and assistance were then requested from the Ministry of Health office for each county before proceeding to the field areas and subsequent engagement with participants via community health workers. Additional research permission granted in Kenya NACOSTI/P/19/43659/

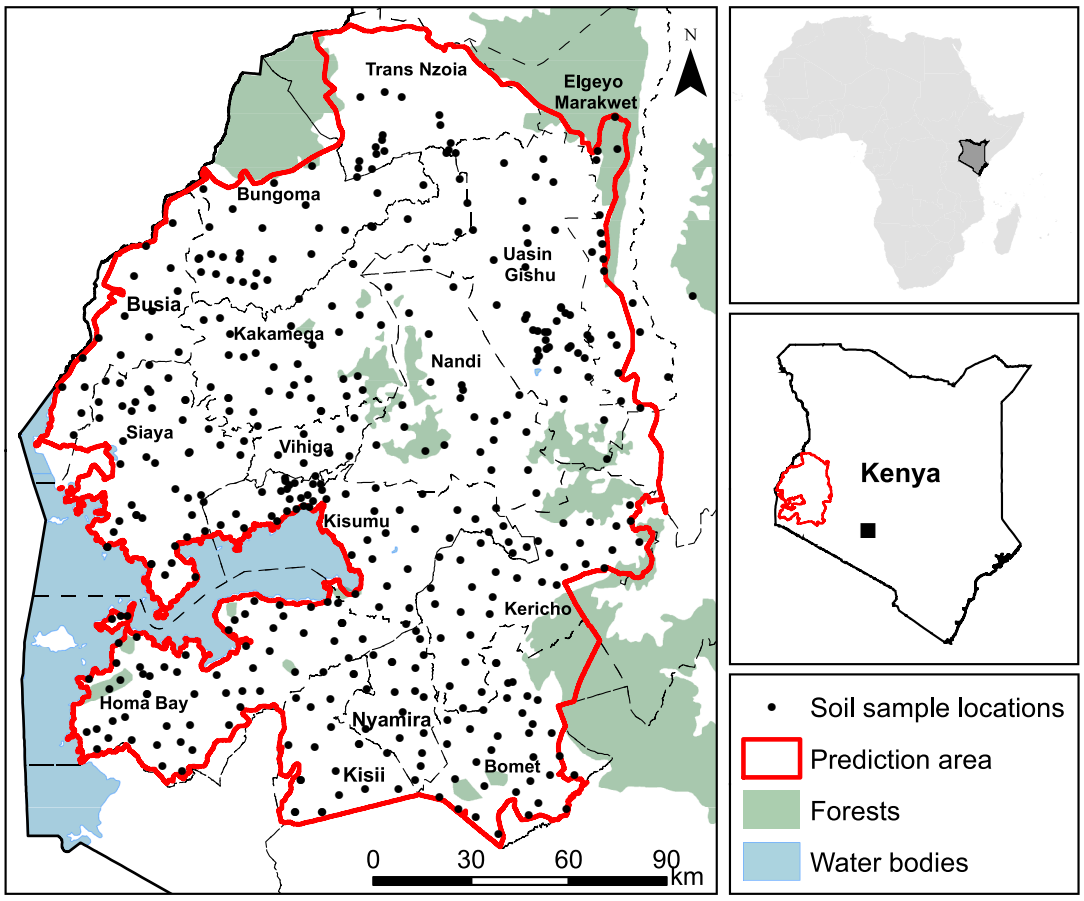


Fig. 1. Location of soil sampling points (n = 446) and prediction area within the study area in western Kenya. Inset maps show the study area in Kenya and Kenya in Africa.

29731.

2.3. Soil data

Soil samples (0.25 g) were digested in a mixed acid solution (HF:2.5 ml/ HNO_3:2 ml/ $HClO_4$:1 ml/ H_2O_2:2.5 ml) for the determination of the total concentrations of a broad suite of major and trace elements on a programmable hot block as described in Watts et al. (2019). For I analyses, soil samples (0.25 g) were digested using 5 ml of 5% *v*/v tetramethylammonium hydroxide (TMAH), heated in a 15 ml Nalgene HDPE bottle at 70 °C in a drying oven for 3 h and then diluted with 5 ml of Milli-Q water followed by centrifugation at 3000 rpm for 20 min, from which the supernatant was used for analyses (Humphrey et al., 2020; Watts and Mitchell, 2009).

The subsequent analyses of the acid digests were performed by an Agilent 8900 triple quadrupole ICP-MS (ICP-QQQ) using (i) collision cell mode (He-gas) for Li, Be, B, Na, Mg, Al, P, S, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, As, Rb, Sr, Y, Zr, Nb, Mo, Ag, Cd, Sn, Sb, Cs, Ba, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Tl, Pb, Bi, Th and U; and (ii) H_2-reaction cell mode for Se; and (iii) O_2-reaction cell mode for As (mass shifted at mass 91). Internal standards Sc, Ge, Rh, In, Te and Ir were employed to correct for signal drift. For I analyses, the ICP-QQQ was operated in 'no-gas' mode and analysed separately as described in Humphrey et al. (2019), with all solutions analysed in a 0.5% TMAH matrix. The complete dataset, limits of detection and analytical performance data are presented in full for certified reference materials in Watts et al. (2021a).

Soil pH analysis was based on a US EPA SW-846 Test Method 9045D for calcareous soils using 5 g of soil (<2 mm), stirred and mixed with a calcium chloride slurry ($CaCl_2$) to a final ratio of 1:2.5. Organic matter content was estimated by loss-on-ignition (LOI) at 450 °C for 1 g of soil, using a < 53 μm particle size.

2.4. Spatial modelling

In this study, a RF modelling approach was adopted where IDW covariates are used as predictor variables. To predict a value for an undetermined location, IDW computes the score of query points based on the scores of their k-nearest neighbours, weighted by the inverse of their distances. As each query point is evaluated using the same number of data points, this method allows for strong gradient changes in regions of high sample density while imposing smoothness in data sparse regions (Fortin and Dale, 2006; Yu and Liu, 2021). Fundamentally, IDW requires 2 parameters the inverse distance power (p) and the number of nearest neighbours (n) to use. The choice of p and n is subjective and there is no relation between the prediction and the actual spatial variability unlike other methods such as kriging where the model relates to the spatial variance of the parameter of interest through a variogram (Wragg and Cave, 2021). In this instance, we have taken a series of IDW predictors with multiple combinations of p and n and use these as the predictor covariates. The RF model combines these covariates to model the estimated soil parameter at the prediction locations.

Geological and geographic features have also been shown to be an important control on the geochemical composition of surface soils (Rawlins et al., 2012). In addition to the IDW predictors, environmental covariates were incorporated into the RF model. These ancillary variables included: elevation (m); average annual rainfall (mm); dominant parent material; major landform; and Euclidean distance to major rivers, urban areas, and waterbodies (Fig. S1, Table S1).

2.5. Random Forest model parameterisation and optimisation

The data analysis was carried out using the R programming Language (R Core Team, 2020) and associated libraries. The "sf" library (Pebesma, 2018) was used to attribute the ancillary data to the sampling points and prediction grid. The "ranger" library (Wright and Ziegler, 2017) was used to carry out the RF modelling and the "Boruta" library (Kursa and Rudnicki, 2010) was used to select the significant predictor variables in the RF model. The Boruta package generates a duplicate predictor (shadow attribute) data set with each predictor randomly shuffled and joins the shuffled dataset with the original predictors. It then builds a RF model on the combined dataset, evaluating the importance of the original variables with the randomised variables. Only variables having higher importance than the randomised variables are considered important and included in the optimised model (Kursa and Rudnicki, 2010). Preliminary checks showed that after the top 5–10 most important IDW predictor combinations, the inclusion of further IDW predictors does not meaningfully improve the root mean square error of the prediction (RMSEP). The modelling and interpolation were performed at 540 m resolution in the following stages:

1. A series of IDW predictor variables were made up from all combinations of n values (3, 5, 7, 9,11,13,15) and *p* values (0.1, 0.5, 0.9, 1.3, 1.7, 2.1, 2.5, 2.9), totalling 56 combinations. For the training set, the IDW predictors are calculated for each individual point using a leave-one-out strategy. A RF model was set up using the 56 IDW combinations as predictor variables for the determinand in question and the performance of the model assessed;
2. The top 10 most important IDW combinations, as assessed in the RF model by the gini-index (Wright and Ziegler, 2017), were chosen and combined with the ancillary variable data. The model was then subjected to the Boruta algorithm which identifies the significant predictors against the shadow attributes (Kursa and Rudnicki, 2010);
3. The RF model was optimised to get the best value of "mtry" (range 1:55), which is the number of variables randomly sampled as candidates at each split in the decision trees used in the RF model (Wright and Ziegler, 2017);
4. Finally, the optimised RF model was applied to 100 bootstrap resamplings of the original sampling points (recalculating the IDW predictors for each bootstrap resample) with each resampling producing data on the model fit and predictions for the element of interest. In the bootstrap sampling method, approximately about two thirds of the samples are used for training the trees, and the remaining one third are used to obtain the out-of-bag (OOB) error which was applied to estimate the performance of the model and guarantee its robustness (Tan et al., 2021). The number of trees selected for all RF models was 1000. The final prediction values were calculated as the median value from all resampling rounds. A prediction error map, calculated as the median absolute deviation of all 100 bootstrap resamplings provides the associated prediction uncertainty.

2.6. Evaluation of algorithm performance and accuracy assessment

The "DALEX" package (Biecek, 2018) was enlisted as a method for post-analysing the RF model to derive the relative importance of the significant predictor variables defined by the Boruta algorithm by measuring how much the root mean square error (RMSE) increases when a given parameter is randomly shuffled, thus determining its importance. To evaluate the performance of the RF model, "out-of-bag" (OOB) predicted values were compared against the measured values of I, Se, Zn and pH in the soil samples used to set up the model. Model performance is ideally assessed using a large independent test dataset that was not used in the training procedure. When an independent dataset is unavailable, k-fold cross-validation is often used, however, RF uses an extension of cross-validation in the form of OOB samples. The OOB data are, in effect, independent samples with measured values not used in setting up the model as within each of the RF model decision trees, bootstrapped samples of the original data were used so that the samples left out by the resampling could act as independent checks. The OOB error provides an estimate of prediction accuracy which is similar to k-fold cross-validation (CV), whilst is may provide a biased estimate of model performance, previous studies have shown that it provides comparable values (Grimm et al., 2008; Svetnik et al., 2003; Hastie et al., 2009; Heung et al., 2014). Moreover, OOB estimates of error are computationally less expensive than k-fold cross-validation.

Ordinary kriging (OK), a geostatistical model, was also used as a benchmark to evaluate the performance of the RF models. A full description of the OK method used was published by Gashu et al. (2021). In brief, summary statistics were calculated and assessed to examine the need to transform the data (Table S2). A decision to use the absolute value or $\log_e$ transformed data was determined based on the coefficient of octile skewness as a robust measure of asymmetry of the distribution (Brys et al., 2004). Variograms were estimated for each variable using three estimators: Matheron (1962), Cressie and Hawkins (1980) and Dowd (1984) (Table S3). A maximum lag distance of 100 km and lag bins with a width of 10 km were used. Exponential variogram functions were fitted to the estimates by weighted least squares (Webster and Oliver, 2007). The model was then tested by cross-validation, the selected models are shown in Fig. S2. The CV process consisted of removing each observation and predicting the remaining observations by ordinary kriging for each variogram model, before assessing the standardised squared prediction error (SSPE) with an expected value of 0.455 (Lark, 2000). The Matheron estimator was selected if the fitted models' CV results were deemed appropriate (the median SSPE is within the 95% confidence interval). If the results suggest the model is not suitable either the Cressie and Hawkins and Dowd estimators were used based on the CV results. The best suited variogram model was then used to compute the predictions of soil concentrations and the kriging variance, as a measure of the uncertainty of the prediction, on a square grid by OK.

The evaluation of algorithm performance, accuracy and quality of the models were assessed using three parameters, root mean squared error (RMSE), Lin's concordance correlation coefficient (CCC), and the coefficient of determination (R^2), with the following formulas:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Pi}-\mathrm{Oi})^2}$$

$$\mathrm{CCC} = \frac{2\,r\,\sigma_o\sigma_p}{\sigma_o^2+\sigma_p^2+(O'-P')^2}$$

$$R^2 = 1-\frac{\sum_i^n(\mathrm{Oi}-\mathrm{Pi})^2}{\sum_{i=1}^n(\mathrm{Oi}-O')^2}$$

where n is the number of samples, Oi and Pi are observed and predicted soil properties, O' and P' are the corresponding means and σ_o^2 and σ_p^2 are the corresponding variances, respectively.

3. Results and discussion

3.1. Iodine

The predicted spatial distribution of total I (mg kg^{-1}) in soil, displayed in deciles, and a measure of the respective uncertainty in western Kenya are shown in Fig. 2. The mean I concentration in this study was 12.05 ± 4.98 mg kg^{-1}, with a range of 2.01–25.43 mg kg^{-1}, and uncertainty estimation between 0.06 and 11.62 mg kg^{-1}. The RF model selected 17 variables for predicting the total concentration of I and optimised the mtry value at 2. The lowest concentrations of I in the study area were located in the northwest of the region and around the

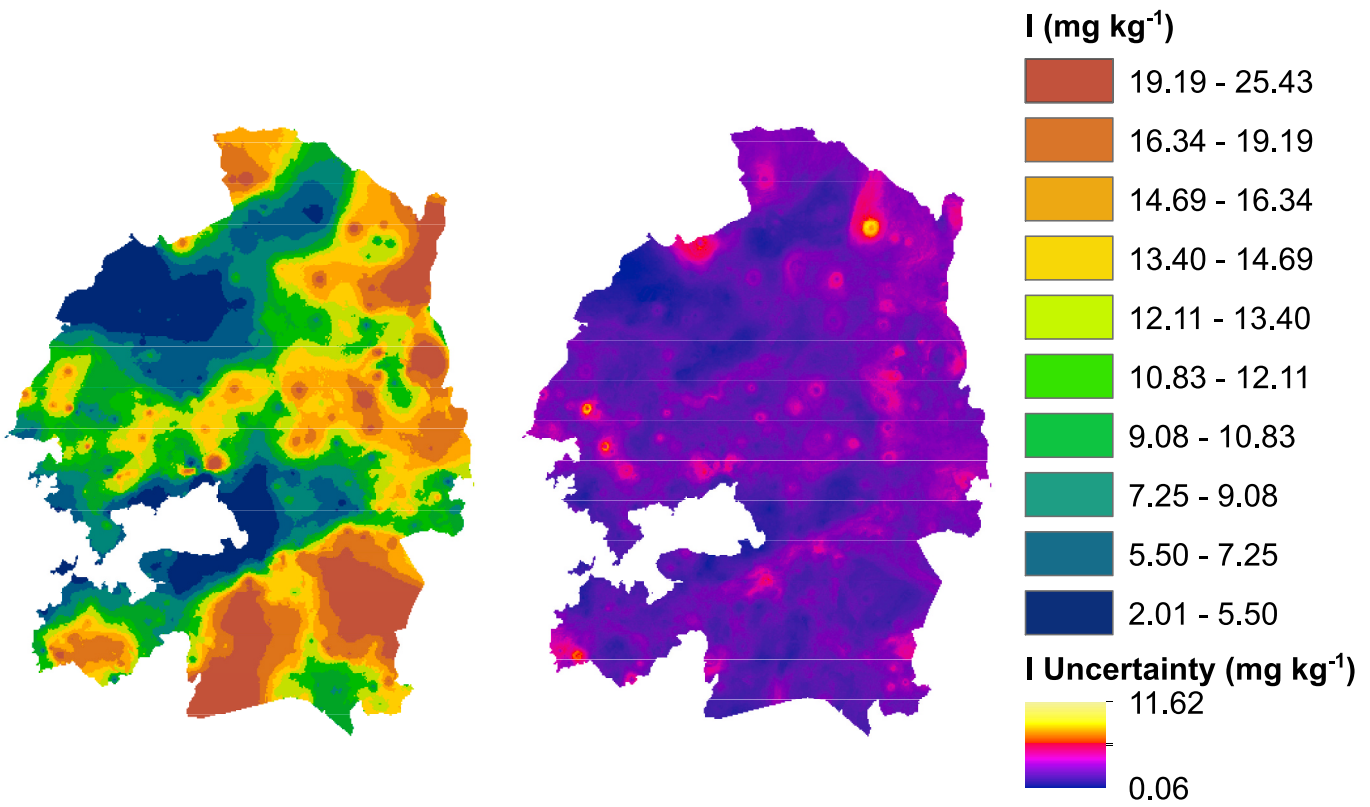


Fig. 2. The spatial prediction of total soil iodine (mg kg^{-1}), displayed in deciles, and prediction uncertainty assessment in western Kenya.

perimeter of the Winam Gulf, whilst the highest concentrations are found in the northeast and southeast. The highest uncertainty was typically limited to localised hotspots between bands of high and low concentrations.

Soil I concentrations are heavily influenced by the wet and dry deposition of volatilised I compounds from marine sources (Humphrey et al., 2018). Despite the large distance of the study area from the coast (>800 km), the average concentration is approximately twice the global soil average (5.1 mg kg^{-1}) (Humphrey et al., 2018; Johnson, 2003). The biogeochemical cycling of I in soils is controlled by soil characteristics that affect retention against leaching, such as soil texture, pH, the concentration of organic matter and metal oxides (Fuge and Johnson, 2015). Ferric and aluminium oxides strongly adsorb I in soils whilst in the presence of manganese oxide birnessite (δ-MnO_2) the inorganic species iodide is oxidised to iodate, however, in the presence of organic matter the intermediate product of the oxidation is incorporated into organic matter (pH < 7) (Humphrey et al., 2020). The soils in western Kenya exhibit a positive Pearson correlation coefficient between I and the total concentration of Fe ($r = 0.57$, $p < 0.01$), Al ($r = 0.63$, $p < 0.01$) and Mn ($r = 0.37$, $p < 0.01$). The soils with the lowest total I (mg kg^{-1}) in the northwest of the study area and surrounding the Winam Gulf have high sand percentages; the highest I concentrations in the southeast region of the prediction area coincide with a landcover dominated by forests with a high organic matter content. Within the study area a positive correlation was found between soil I and organic matter content ($r = 0.44$, $p < 0.01$). These observations agree with Johnson (2003), who identified that soil texture affects I enrichment in soils and that soils with high sand ratios have lower I concentrations, whereas organic-rich soils can retain higher I concentrations (Humphrey et al., 2020). Rickard and Price (1984) reported that soils at high elevations have greater concentrations of I than low elevations soils. Within the prediction area there is good general agreement between I and elevation ($r = 0.47$, $p < 0.01$). Orographic lifting and subsequent precipitation may have a major influence on I concentrations in soils in western Kenya and has previously been identified as an important factor in determining I concentrations in terrestrial precipitation (Gilfedder et al., 2007).

The relative importance of all predictor variables and the partial dependence of the top 10 predictor variables for I are shown in Fig. 3. Whilst the relative importance shows how integral the predictor variable is to the overall prediction, the partial dependence plot shows the marginal effect a single feature has on the predicted outcome of a machine learning model and can indicate whether the relationship between the target and a feature is linear, monotonic or more complex (Friedman, 2001). The variable importance plot of the RF model revealed that the IDW predictor variables have the greatest impact on predicting the distribution of total I in the study area, with a combined relative importance of 60.7%. The most important ancillary variable was elevation, followed by distance to waterbodies, rainfall and proximity to urban areas, respectively, these variables have a combined relative importance of 24.3%. The remaining geological and landform variables combine to have a 15% importance. The limited importance of these features in comparison to the IDW variables, elevation and rainfall highlight the unique biogeochemical cycling on I in the terrestrial environment. All of the IDW partial dependence plots follow a similar profile. The partial dependence plot for elevation shows a sharp increase between 1800 and 2100 m, prior to plateauing supporting the notion that orographic lifting and subsequent precipitation may significantly impact soil I concentrations. Interestingly, the partial dependence plot for distance to waterbodies shows that the closeness to a waterbody negatively affects the concentration of I in the soil. The largest waterbody in this study area is the Winam Gulf, and soils close to the gulf are sandy, thereby supporting previous observations on the importance of soil texture on I concentrations.

3.2. Selenium

The predicted spatial distribution of total Se (mg kg^{-1}) in soil, displayed in deciles, and a measure of the respective uncertainty in western

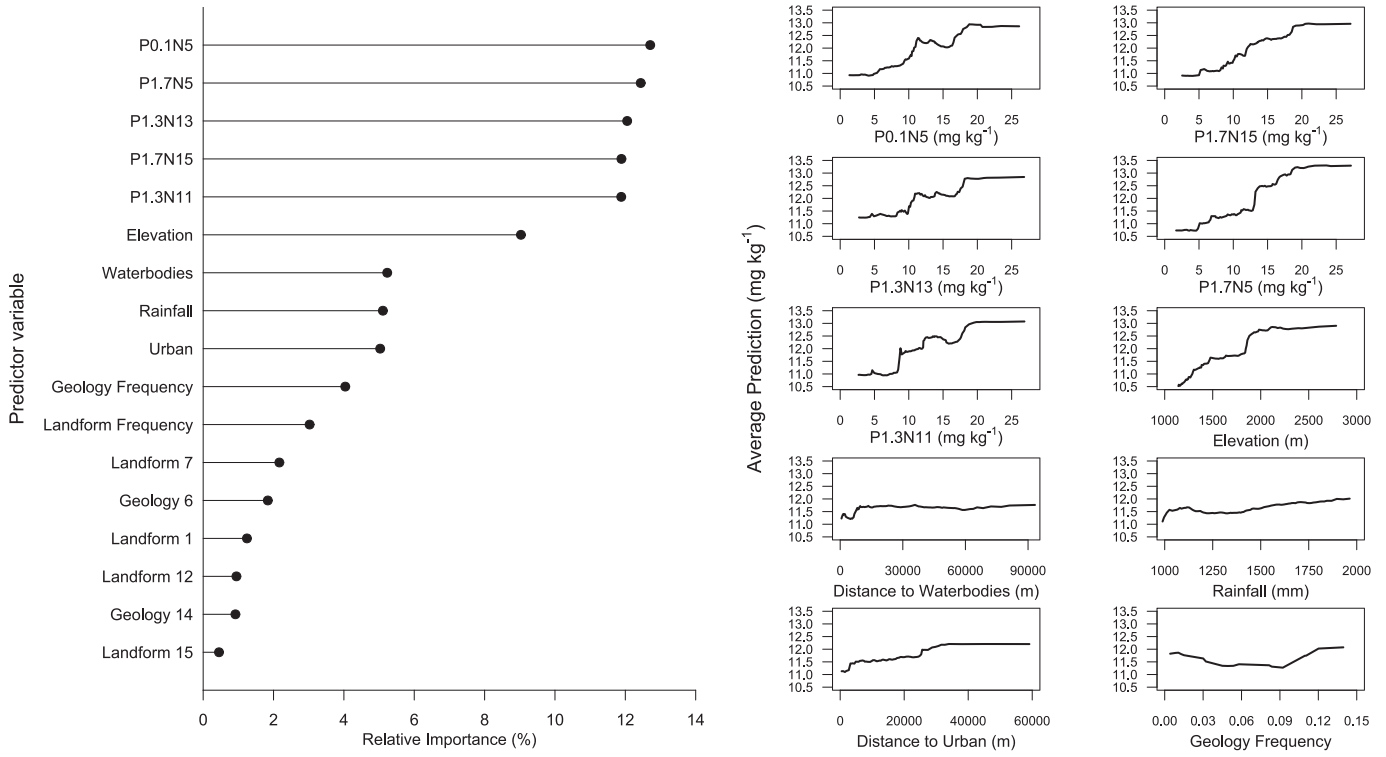


Fig. 3. Scaled variable importance plot from random forest model for factors predicting total iodine concentration (mg kg^{-1}) and response curves illustrating the relationship between iodine concentration (mg kg^{-1}) and the input variables in soils within the study area.

Kenya are shown in Fig. 4. The predicted Se concentrations in the study area range from 0.24 to 1.52 mg kg^{-1}, which matches the global range of 0.01–2.00 mg kg^{-1} (Dungan and Frankenberger, 1999). The predicted mean Se concentration in the study area (0.71 ± 0.29 mg kg^{-1}) is slightly elevated in comparison to the global average (0.4 mg kg^{-1}). The RF model selected 22 variables for predicting the total concentration of Se and optimised the mtry value at 2.

Selenium is an essential trace element for human health with foodstuffs being an important source of Se to humans; however, Se intake is highly variable due to its heterogeneous distribution in the terrestrial environment and agricultural soils (Rayman, 2012). Whilst bedrock geology appears to be the most important factor for Se-rich soils it fails to explain the large-scale distribution of Se (Blazina et al., 2014). The ocean is an important reservoir of Se from where it can be transported inland via volatilisation and deposition events (Blazina et al., 2014). The spatial distribution of Se in the study area follows a similar pattern to I with the lowest concentrations in the northwest of the region and the area surrounding the Winam Gulf, whilst the highest concentrations are found in forested areas to the southeast. Haibo et al. (2005) reported similar concentrations of Se in woodland (1.36 mg kg^{-1}), grassland (0.67 mg kg^{-1}) and arable cropland (0.36 mg kg^{-1}) in Hong Kong. The comparable observations indicate that the concentrations of Se in soils vary with different landcover. In the present study, the highest uncertainty was found in the central region of the study area and between bands of high and low concentrations. The dominant species of Se in soils are oxyanions selenite and selenate in acidic and alkaline soil, respectively, which are known to form complexes with clay minerals, organic matter, and adsorb to oxyhydroxides (Blazina et al., 2014; Xu et al., 2020). In the current study a positive relationship was found between Se and the total concentration of Fe ($r = 0.55$, $p < 0.01$), Al ($r = 0.57$, $p < 0.01$) and Mn ($r = 0.49$, $p < 0.01$), suggesting that soils with high content in oxyhydroxides can adsorb and retain more Se. However, the total concentration of Se does not accurately portray crop, and subsequent dietary availability as the mobility of Se in soil is affected by various biological and geochemical factors such as redox potential, pH, and soil organic matter (Pisarek et al., 2021). In Malawi, Hurst et al. (2013) observed that on calcareous soils (eutric vertisols), soil-to-crop transfer of Se was >10-fold higher compared to other soils and that pH markedly affected dietary Se intake. There was a negative correlation between Se and soil pH ($r = -0.33$, $p < 0.01$) in this study, supporting previously published observations (Haibo et al., 2005). Due to the complexity of Se soil dynamics, additional research is required to fully utilise this predictive map.

The relative importance of all predictor variables and the partial dependence of the top 10 predictor variables for Se are shown in Fig. 5. The variable importance plot of the RF model shows that the IDW predictor variables have the greatest impact on predicting the distribution of total Se in the study area, with a combined relative importance of 62.4%. The most important ancillary variable was elevation (8%), followed by distance to waterbodies (4.4%), rainfall (4.3%), and landform frequency (3.7%). The remaining predictor variables (geological features) have a combined relative importance of 17.2%. In this study, there is a strong positive correlation ($r = 0.77$, $p < 0.01$) between I and Se, as expected due to the similarity of factors influencing their biogeochemical cycling and the similarities between the partial dependence plots. The partial dependence plots show a positive relationship for both elevation and rainfall with the total concentrations of Se in the soils of western Kenya. Shao et al. (2018) observed that soil Se concentrations decreased from upstream to downstream in a watershed in southern China, which significantly correlated with elevation. In the present study a positive correlation ($r = 0.49$, $p < 0.01$) was also found between Se and elevation. Blazina et al. (2014) observed that high soil Se concentrations in northwest China are likely due to enrichment in drier saline-alkaline soils. However, the similarity between Se soil distribution and precipitation indicates that atmospheric Se inputs via precipitation also play an important role, and could be influential in other regions worldwide. The results in this study highlight that rainfall plays an important role in the biogeochemical cycling of Se.

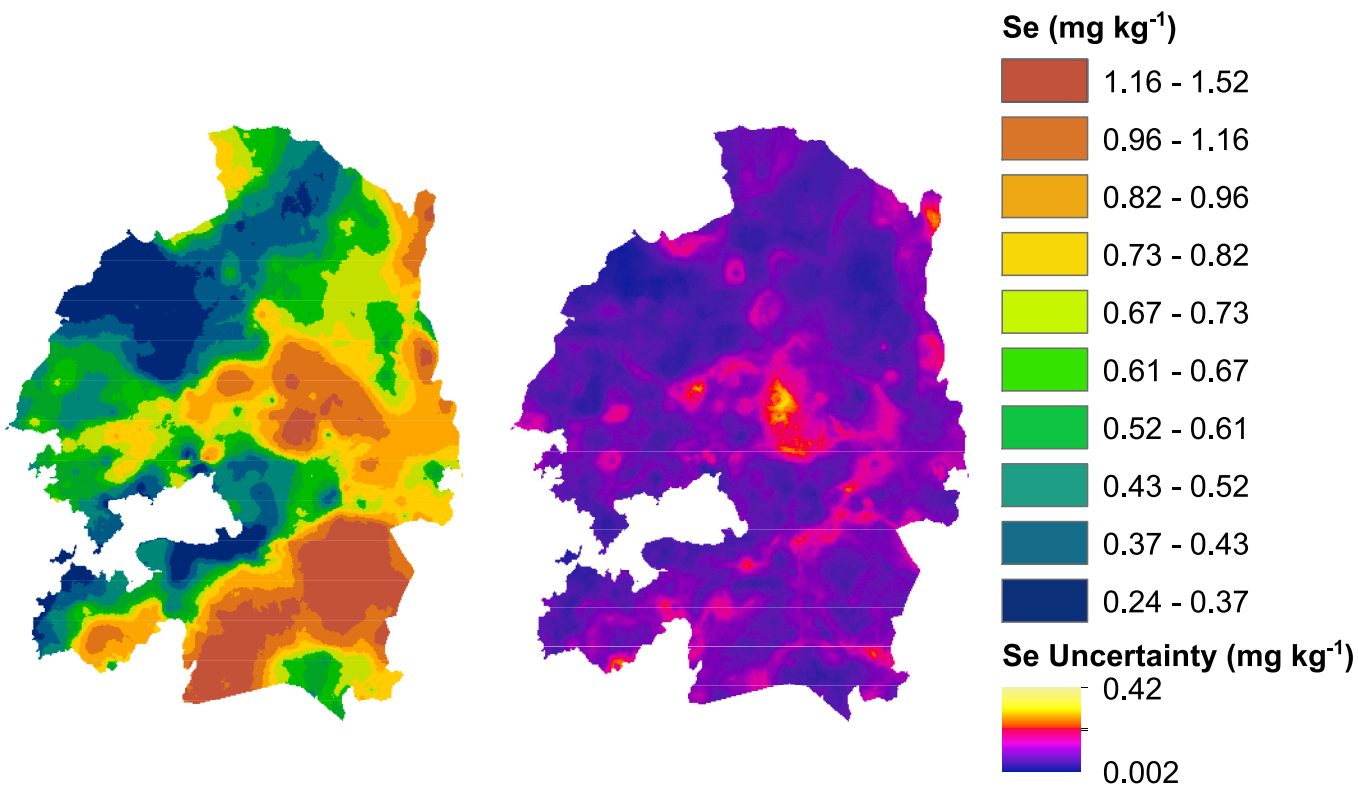


Fig. 4. The spatial prediction of total soil selenium (mg kg^{-1}), displayed in deciles, and prediction uncertainty assessment in western Kenya.

3.3. Zinc

The predicted spatial distribution of total Zn (mg kg^{-1}) in soil, displayed in deciles, and an assessment of prediction uncertainty in western Kenya are shown in Fig. 6. The mean Zn concentration in the study area is 147.1 ± 67.8 mg kg^{-1}, with a range of 36.6–366.9 mg kg^{-1}, and an uncertainty estimation between 0.3 and 140.0 mg kg^{-1}. The average concentration of Zn in the current study is higher than the global average (55.0 mg kg^{-1}), however, our results are comparable to the typical range of Zn reported in soils (10.0–300.0 mg kg^{-1}) by Alloway (1995). The spatial distribution of Zn in the study area generally increases from northwest to southeast. The lowest concentrations (36.6–62.5 mg kg^{-1}) are located in the northwest and the south-eastern shore of the Winam Gulf, whilst the highest concentrations (230.9–366.9 mg kg^{-1}) are found in a central hotspot and to the south-east of the study area. The highest uncertainty was found at a Zn hotspot located in the middle of the prediction area located in the middle of the urban centre of Kisumu, which is likely to explain the high concentrations of Zn. The RF model selected 17 variables for predicting the total concentration of Zn and optimised the mtry value at 2.

Zinc is an essential micronutrient for humans and plants and has diverse physiological functions in biological systems. It interacts with a large number of enzymes and other proteins in the body and performs critical structural, functional and regulatory roles (Cakmak and Kutman, 2018). The total Zn content of a soil is largely dependent upon the geochemical composition of the weathering rock parent material on which the soil has developed, however, Zn can be incorporated into soils through anthropogenic activities such as pollution or fertiliser application (Alloway, 2008). Typically, high concentrations of Zn in soil correlate with basic igneous rocks, such as basalts due to Zn occurring in ferromagnesian minerals, where it has isomorphously substituted Fe^{2+} or Mg^{2+} (Alloway, 2008). Conversely, silica-rich igneous rocks have lower total Zn contents and their residual weathering product is usually quartz sand which gives rise to sandy soils with low concentrations of Zn and other essential micronutrients (Alloway, 2008). The results in the present study support these findings with the highest concentrations of Zn overlaying volcanic and basic igneous rocks in the southeast of the study site, whilst soils with the lowest concentrations have formed on quartz-rich granites and gneiss parent material (Fig. S1-F). Furthermore, soil texture plays an important role in the total concentration of Zn found in soils with clay-rich soils having greater capacity to adsorb and retain Zn relative to soils with lower percentages of clay and higher percentages of sand (Baize, 1997; Gorny et al., 2000; McGrath and Loveland, 1992). Despite the relative abundance of Zn in the soil of the study area Watts et al. (2019) calculated that the risk of Zn dietary deficiency, based on biomonitoring data, exceeded 90% in this region. In agricultural soils Zn is bound to clays, hydrous oxides and organic material depending on various physicochemical soil factors mainly pH and organic matter content (Noulas et al., 2018). These factors ultimately determine the solubility of Zn in soil, and consequently, its bioavailability for uptake by plants. As such, further research is required with additional layers to create predictive maps capable of providing recommendations to farmers for Zn fertiliser applications.

The relative importance of all predictor variables and the partial dependence of the top 10 predictor variables for Zn are shown in Fig. 7. The variable importance plot of the RF model shows that the IDW

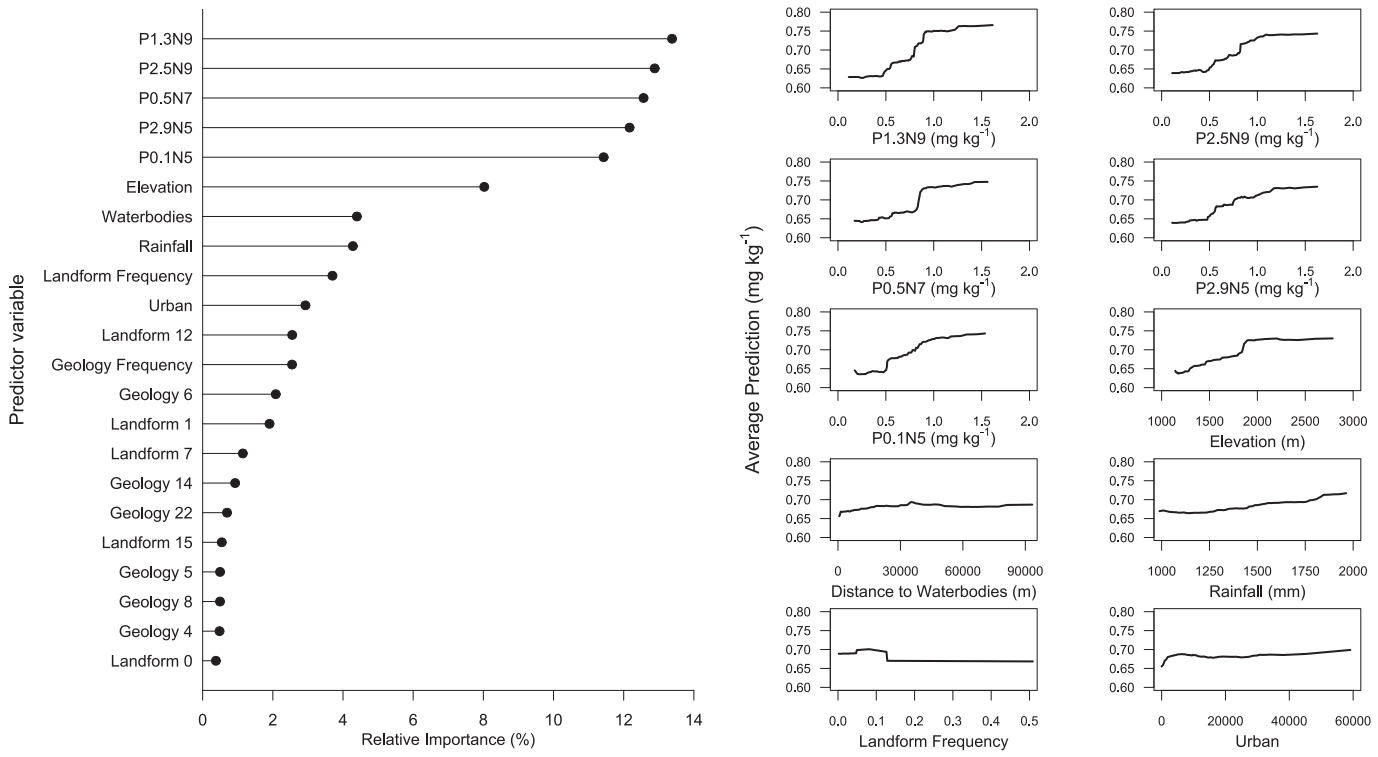


Fig. 5. Scaled variable importance plot from random forest model for factors predicting total selenium concentration (mg kg^{-1}) and response curves illustrating the relationship between selenium concentration (mg kg^{-1}) and the input variables in soils within the study area.

predictor variables have the greatest impact for assessing the distribution of total Zn in the study area, with a combined relative importance of 71.4%. In addition, the IDW variables all display a similar partial dependence profile. The most important ancillary variables were elevation (5.8%), waterbodies (4.8%), rainfall (4.2%), and the remaining 13.8% was attributed to geological/landform features. The comparatively greater importance of the combined geological and landform features suggests that soil Zn concentrations were heavily influenced by the underlying geology.

3.4. Soil pH

The predicted spatial distribution of soil pH, displayed in deciles, and a correspondent measure of uncertainty in western Kenya are shown in Fig. 8. The mean soil pH in this region is 5.59 ± 0.47, with a range of 4.45–7.56 and an uncertainty estimation between 0.01 and 1.08. The RF model selected fewer significant independent variables (10) for predicting soil pH in comparison to I, Se and Zn models and optimised the mtry value to 8, which is greater compared to the other models presented in this paper. The acidic nature of soils in western Kenya is well documented (Kisinyo et al., 2014). In the present study, the most acidic soils (4.46–5.09) were found in the northeast of the prediction area. Only 12% of all the soils collected in this study had a soil pH >7 which were predominantly located around the Winam Gulf and to the west of the study site.

Soil pH is considered to be the "master variable" of soil chemistry due to its profound impact on chemical reactions involving essential nutrients and potentially toxic elements, therefore soil pH management is critical for both agronomic and environmental management (Neina, 2019; Penn and Camberato, 2019). Inherent factors that affect soil pH include climate, mineralogy, and topography. The pH of a young soil is typically determined by the mineralogy of its parent material, whilst in older soils temperature and rainfall affect the intensity of leaching and the weathering of soil minerals. In warm, humid environments, soil pH decreases over time through acidification due to leaching caused by the high volume of rainfall (Fabian et al., 2014; Li et al., 2017; Slessarev et al., 2016). It has previously been reported that topographic factors are significantly correlated with soil pH at a range of latitudes (Chen et al., 1997; Seibert et al., 2007). Within the current study, there is a significant negative trend between soil pH and elevation ($r = -0.44$, $p < 0.01$), with more acidic soils present at higher altitudes, subjected to high organic matter contents and greater precipitation, to the east of the study site (Fig. 8). Approximately 13% of Kenyan soils are acidic, with the majority found in western Kenya and the Rift valley which developed from non-calcareous parent materials such as syenites, phonolites, trachytes, olivines, older basic tuffs and nepholites (Kisinyo et al., 2014). The predicted results presented in the current study have good agreement with previously published values in the same region (George et al., 2002; Moebius-Clune et al., 2011; Muindi et al., 2015; Opala et al., 2018; Otieno and Zingore, 2018). Due to the acidic nature of the soils in western Kenya, the vast majority of them have high exchangeable Al_3^+ ions which have led to high P sorption in these soils significantly reducing P availability to staple crops, such as maize, severely limiting crop yields in the area. Crop production in acid soils with Al toxicity and low soil available P may be improved by the use of lime and/or fertilisers with liming effects, as the application of agricultural lime increases Ca^+/Mg^{2+} ions and reduces the presence of Al^{3+}, H^+, Mn^{4+}, and Fe^{3+} ions in soil solution (Kisinyo et al., 2014). The results presented in the current study could be used to provide agricultural practice recommendations to improve crop yields and improve micronutrient supply to the population in western Kenya.

The relative importance of all predictor variables and the partial dependence of the top 10 predictor variables for soil pH are shown in Fig. 9. The variable importance plot of the RF model shows that the IDW predictor variables have the greatest impact on predicting the distribution of soil pH in the study area, with a combined relative importance of 63.9%. The most important ancillary variable was elevation (11.7%), followed by distance to waterbodies (9.9%), rainfall (9.2%), and geological and landform frequency (5.2%). In contrast to the elemental variable importance plots, elevation and rainfall have a markedly higher

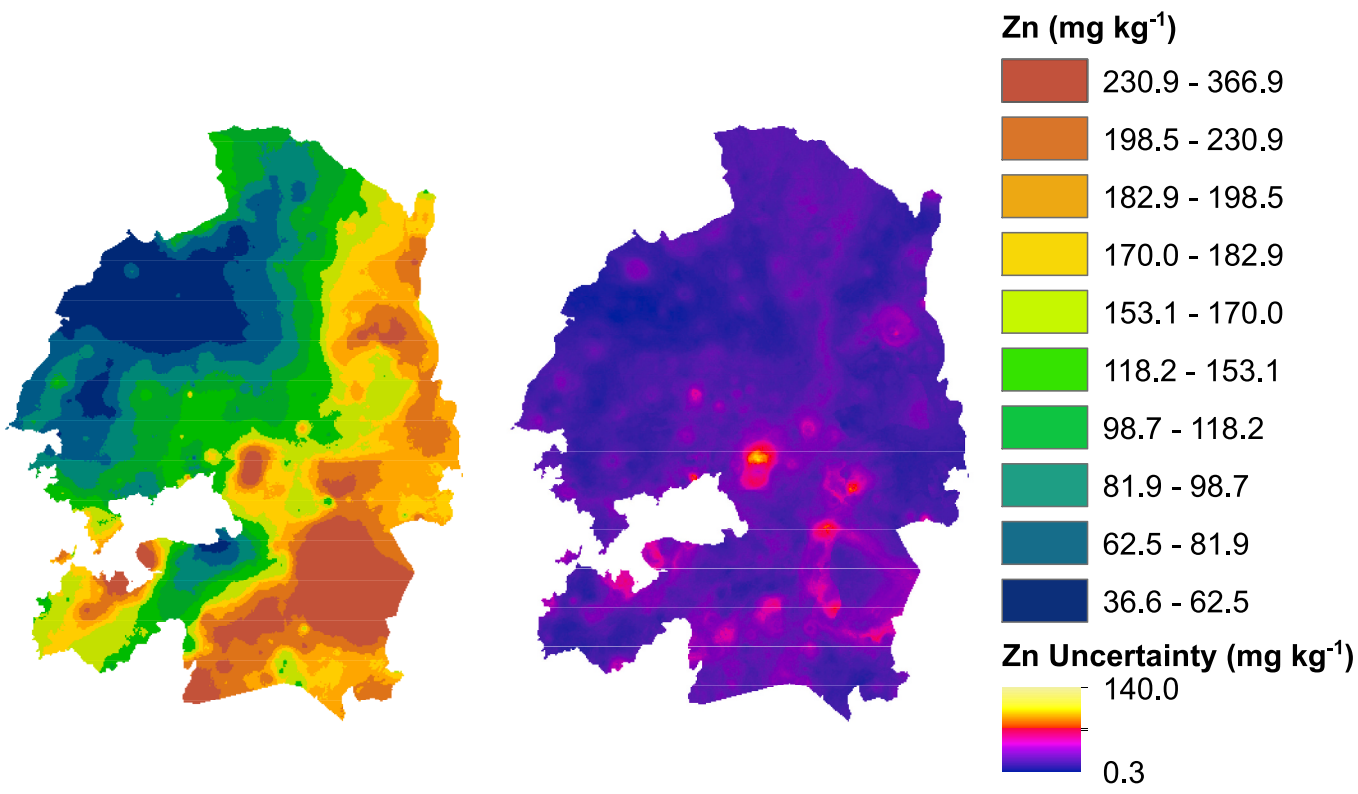


Fig. 6. The spatial prediction of total soil zinc (mg kg^{-1}), displayed in deciles, and prediction uncertainty assessment in western Kenya.

importance in the RF model for predicting soil pH. The response curves for elevation and rainfall (Fig. 9) both indicate that with increased elevation and rainfall soils become more acidic in western Kenya. Elevation significantly influences local precipitation rates. In general, elevation and precipitation are positively correlated, and excess rainfall leads to the leaching of Ca and Mg ions, thereby acidifying soils (Badraghi et al., 2021).

3.5. Model performance

The accuracy, assessed by RMSE, CCC, and R^2 values, of the spatially predicted soil properties where IDW covariates are used in conjunction with ancillary environmental covariates were compared to RF modelling only using IDW covariates and ordinary kriging (Table 1). The RF model with IDW and ancillary covariates and the OK interpolation method outperformed the RF model only using IDW covariates highlighting the importance of including ancillary values in the RF model. Furthermore, the RF model (IDW + Ancillary covariates) described in this paper surpassed or equalled all accuracy assessment indices compared to OK; excluding Zn RMSE where OK performed marginally better.

To further evaluate the performance of the RF model (IDW + Ancillary covariates) OOB predicted values were compared against the measured values of I, Se, Zn and pH in the soil samples used to set up the model (Fig. S3). The OOB data act as independent samples with measured values not used in setting up the model as within each of the RF model decision trees, bootstrapped samples of the original data were used so that the samples omitted by the resampling could act as independent checks (Breiman, 2001). The distance of the points from the line of equivalence and the size of the vertical error bars on the OOB predictions provide the estimates for the accuracy and precision, respectively. In general, the OOB-predicted values agreed with the actual values within the errors indicated by the error bars for all soil properties. However, it appears that there is a negative bias for the highest concentrations and a positive bias for the lowest concentrations for all soil properties. Despite the large variation and complex heterogeneity of total elemental concentrations and soil pH within the study area, which can be attributed to differences in parent materials between and within sites and to local pedologic, hydrological, and management factors our RF model using both IDW and ancillary covariates models performed very well.

4. Conclusion

In this study, we applied RF modelling utilising both IDW and ancillary covariates for spatially predicting and assessing the uncertainty of the total concentration (mg kg^{-1}) of I, Se, Zn and soil pH in western Kenya. The RF models were optimised using a feature ranking algorithm identifying statistically significant variables and a powerful non-parametric post-processing tool was used to explain the importance of significant attributes used in the model. Whilst the IDW predictors were the most important covariates in all models the inclusion of ancillary covariates improved the prediction capability. Moreover, the RF models used largely outperformed ordinary kriging based on multiple accuracy assessment indices (RMSE, CCC and R^2). The RF modelling used in this paper where ML methods are used in conjunction with readily available environmental datasets could provide significant

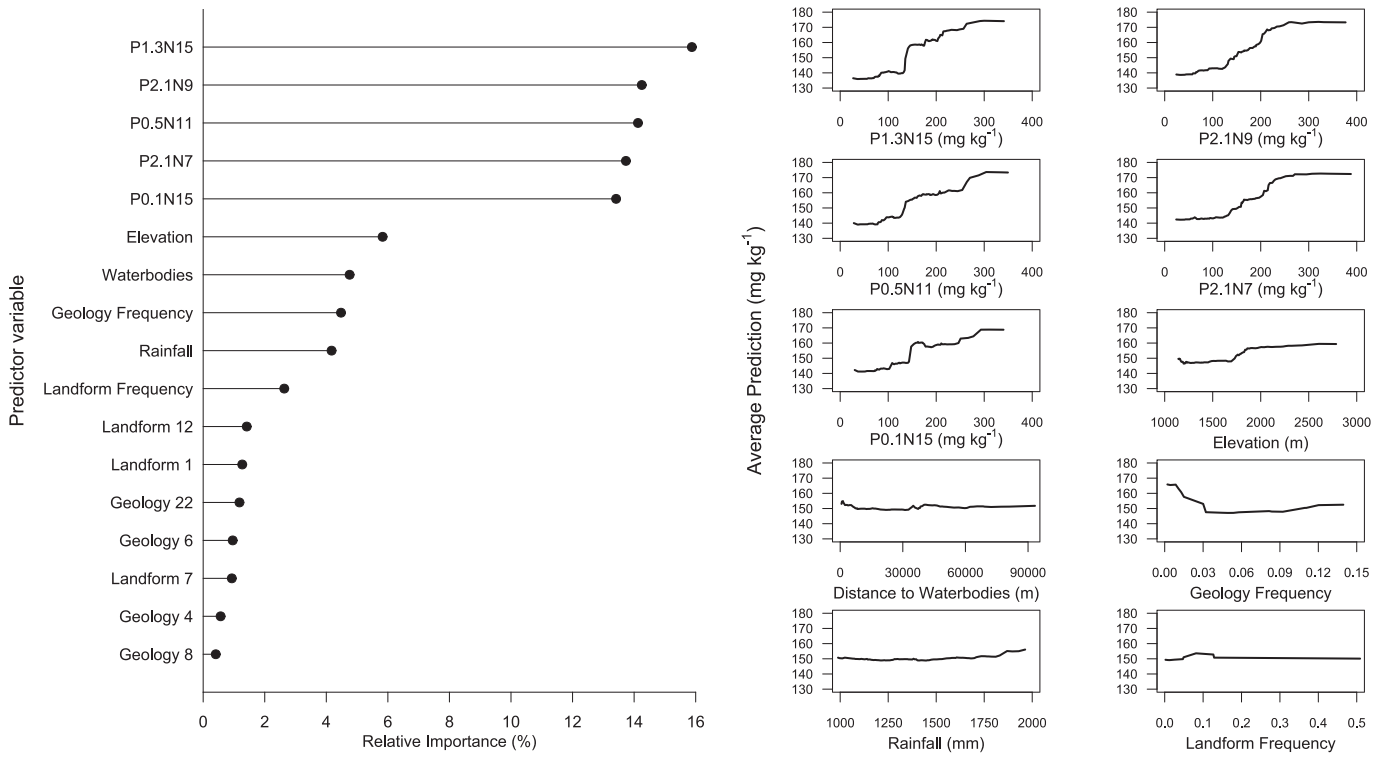


Fig. 7. Scaled variable importance plot from random forest model for factors predicting total zinc concentration (mg kg^{-1}) and response curves illustrating the relationship between zinc concentration (mg kg^{-1}) and the input variables in soils within the study area.

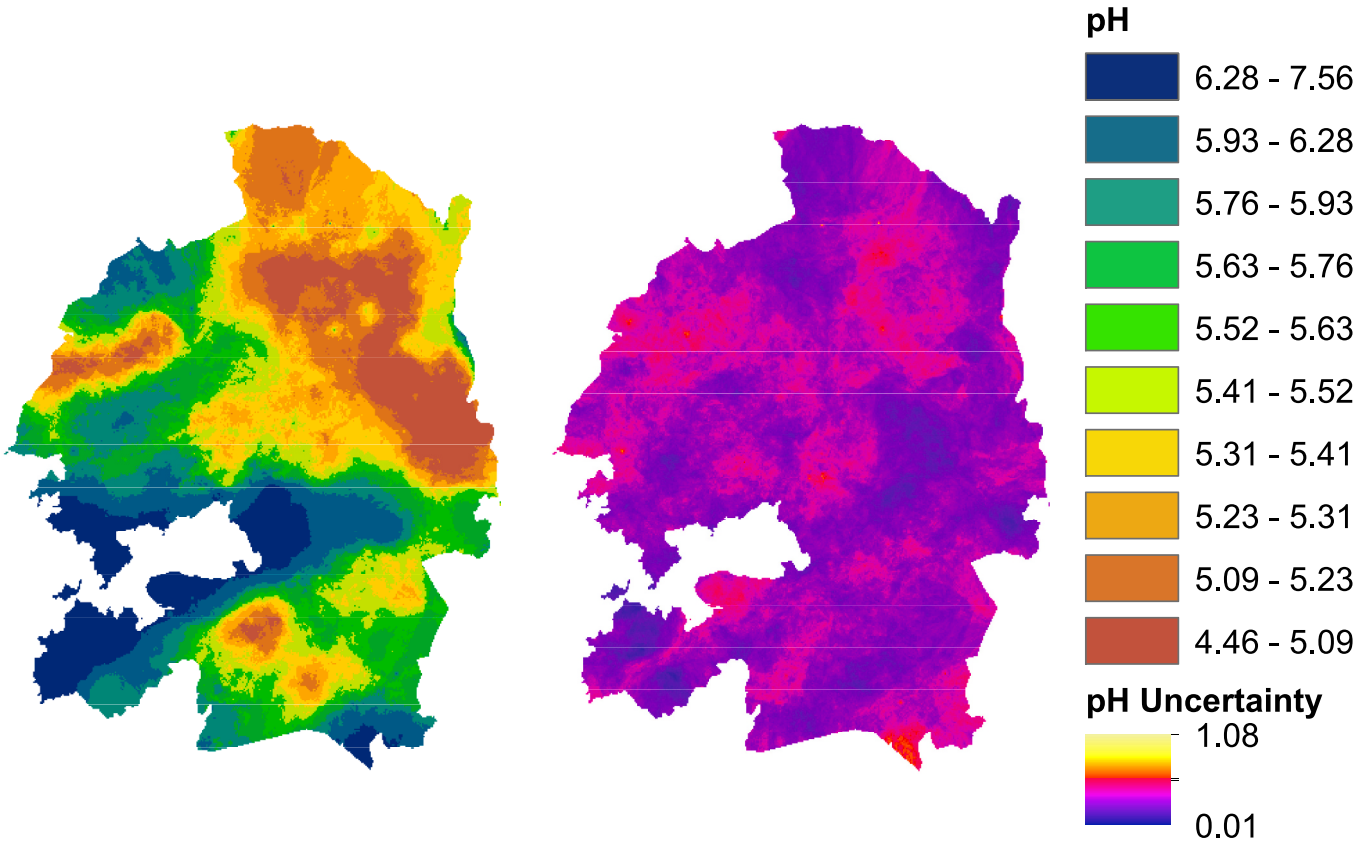


Fig. 8. The spatial prediction of soil pH, displayed in deciles, and prediction uncertainty assessment in western Kenya.

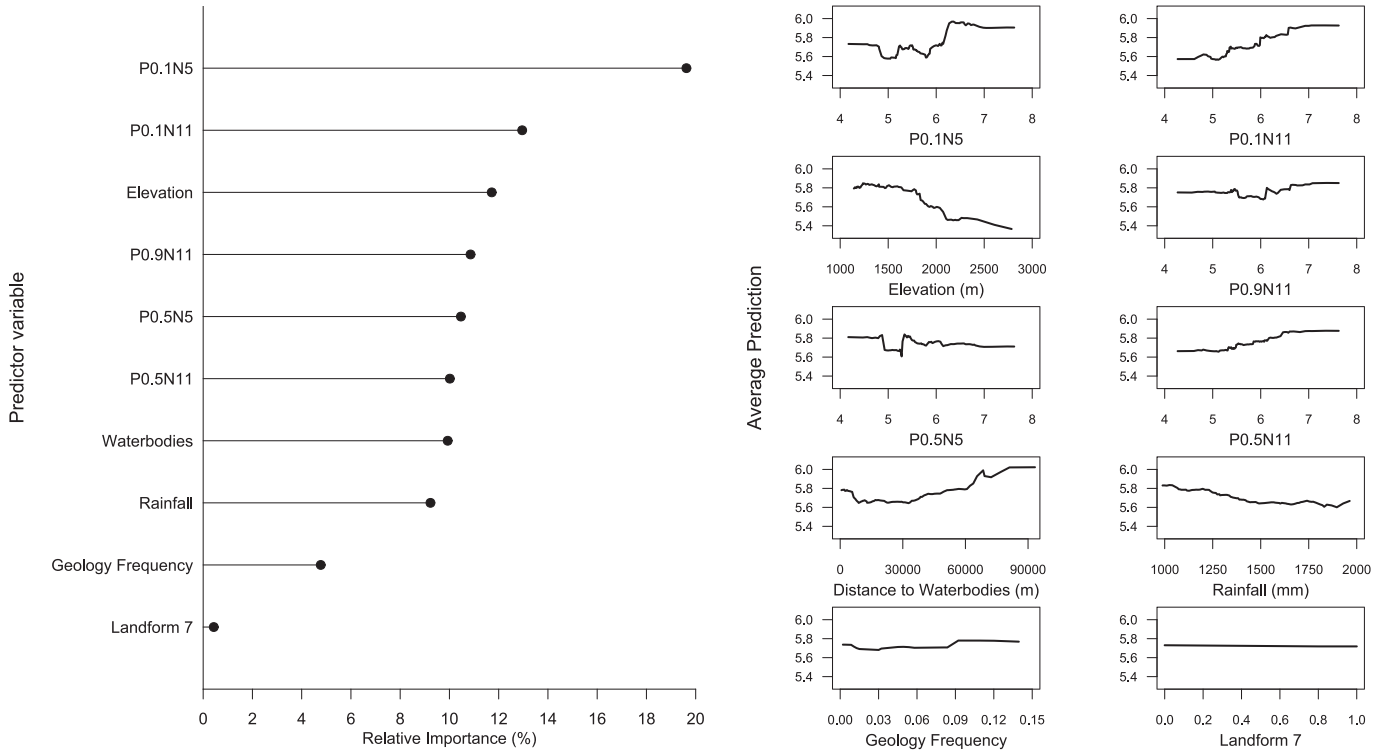


Fig. 9. Scaled variable importance plot from random forest model for factors predicting soil pH and response curves illustrating the relationship between soil pH and the input variables in soils within the study area.

Table 1
Comparison of model performance assessment indices for iodine, selenium, zinc and soil pH using different interpolation methods. The best performing model assessment indices are shown in italics.

Soil property	Accuracy assessment indices	Interpolation method		
		RF (IDW + Ancillary covariates)[a]	RF (IDW covariates)	OK
I (mg kg^{-1})	RMSE	*5.53*	5.92	5.75
	CCC	*0.63*	0.58	0.59
	R^2	*0.46*	0.38	0.42
Se (mg kg^{-1})	RMSE	*0.22*	0.24	0.23
	CCC	*0.77*	0.74	0.76
	R^2	*0.64*	0.58	0.61
Zn (mg kg^{-1})	RMSE	58.6	59.5	*54.4*
	CCC	*0.72*	*0.72*	0.71
	R^2	*0.56*	0.54	*0.56*
pH	RMSE	*0.78*	0.77	*0.78*
	CCC	*0.51*	0.47	0.47
	R^2	*0.31*	0.29	*0.31*

[a] Model used to create spatial prediction maps.

improvements to the accuracy and resolution of soil prediction maps. In addition, the model presented in this paper has been applied to spatially predict the total concentration of 56 elements in western Kenya, the open-access database is available in Watts et al. (2021a). Research is now required to assess how the geochemical maps presented can be used in conjunction with health practitioners to investigate the relationship between environmental geochemistry and endemic diseases, such as esophageal cancer. The method presented in this paper can be used harmoniously with additional layers of soil, staple crop and urinary biomarker data, to define geographic areas that would be at an elevated risk of micronutrient deficiency. Furthermore, the outputs from this modelling framework could be used to guide agricultural policy and farmer recommendations for fertiliser application or soil physiochemical amendment intervention.

CRediT authorship contribution statement

Olivier S. Humphrey: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Mark Cave:** Methodology, Software, Validation, Writing – review & editing. **Elliott M. Hamilton:** Formal analysis, Validation, Visualization, Writing – review & editing. **Odipo Osano:** Conceptualization, Writing – review & editing. **Diana Menya:** Writing – review & editing. **Michael J. Watts:** Conceptualization, Funding acquisition, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgements

Funding for this research was provided by the Royal Society (grant number: ICA\R1\1910770) and NERC-UKRI ODA-NC Foundation Award (grant number: NE/R000069/1). The authors would like to thank all data providers for making their data available. This work is published with the permission of the Executive Director, British Geological Survey.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2023.e00731.

References

Alloway, B., 1995. Heavy Metals in Soils. Blackie Academic and Professional. An Imprint of Chapman & Hall, Glasgow.

Alloway, B.J., 2008. Zinc in Soils and Crop Nutrition.

Asgari, N., Ayoubi, S., Demattê, J.A.M., Jafari, A., Safanelli, J.L., Silveira, A.F.D.D., 2020. Digital mapping of soil drainage using remote sensing, DEM and soil color in a semiarid region of Central Iran. Geoderma Reg. 22, e00302.

Badraghi, A., Ventura, M., Polo, A., Borruso, L., Giammarchi, F., Montagnani, L., 2021. Soil respiration variation along an altitudinal gradient in the Italian Alps: disentangling forest structure and temperature effects. PLoS One 16 (8), e0247893.

Baize, D., 1997. Teneurs totales en éléments traces métalliques dans les sols (France): Références et stratégies d'interprétation. Programme ASPITET. Editions Quae.

Biecek, P., 2018. DALEX: explainers for complex predictive models in R. J. Mach. Learn. Res. 19 (1), 3245–3249.

Blazina, T., Sun, Y., Voegelin, A., Lenz, M., Berg, M., Winkel, L.H.E., 2014. Terrestrial selenium distribution in China is potentially linked to monsoonal climate. Nat. Commun. 5 (1), 4717.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. J. Comput. Graph. Stat. 13 (4), 996–1017.

Cakmak, I., Kutman, U.B., 2018. Agronomic biofortification of cereals with zinc: a review. Eur. J. Soil Sci. 69 (1), 172–180.

Cave, M., 2017. A machine learning approach to Geostatistics applied to contaminants in soil 6D.2. In: 7th International Contaminated Site Remediation Conference Incorporating the 1st International PFAS Conference. CRC CARE 2017, Melbourne, Australia, pp. 330–331.

Chagas, C.D.S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. CATENA 139, 232–240.

Chen, Z.-S., Hsieh, C.-F., Jiang, F.-Y., Hsieh, T.-H., Sun, I.-F., 1997. Relations of soil properties to topography and vegetation in a subtropical rain forest in southern Taiwan. Plant Ecol. 132 (2), 229–241.

Cressie, N., Hawkins, D.M., 1980. Robust estimation of the variogram: I. J. Int. Assoc. Math. Geol. 12 (2), 115–125.

Dowd, P., 1984. The Variogram and Kriging: Robust and Resistant Estimators, Geostatistics for Natural Resources Characterization. Springer, pp. 91–106.

Dungan, R., Frankenberger, W., 1999. Microbial transformations of selenium and the bioremediation of seleniferous environments. Biorem. J. 3 (3), 171–188.

Fabian, C., Reimann, C., Fabian, K., Birke, M., Baritz, R., Haslinger, E., 2014. GEMAS: Spatial distribution of the pH of European agricultural and grazing land soil. Appl. Geochem. 48, 207–216.

Fortin, M.-J., Dale, M., 2006. Spatial Analysis: A guide for Ecologists. Cambridge University Press, Cambridge.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Fuge, R., Johnson, C.C., 2015. Iodine and human health, the role of environmental geochemistry and diet, a review. Appl. Geochem. 63, 282–302.

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. The nutritional quality of cereals varies geospatially in Ethiopia and Malawi. Nature 594 (7861), 71–76.

George, T., Gregory, P., Robinson, J., Buresh, R., Jama, B., 2002. Utilisation of soil organic P by agroforestry and crop species in the field, western Kenya. Plant Soil 246 (1), 53–63.

Gilfedder, B.S., Petri, M., Biester, H., 2007. Iodine and bromine speciation in snow and the effect of orographically induced precipitation. Atmos. Chem. Phys. 7 (10), 2661–2669.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89 (1–2), 1–45.

Gorny, A., Uterman, J., Eckelmann, W., 2000. Germany: in Heavy Metal (Trace Element) and Organic Matter Contents of European Soils, 30. European Comission. CEN Soil Team.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. Geoderma 146 (1–2), 102–113.

Haibo, Z., Yongming, L., Longhua, W., 2005. Hong Kong soil researches II. Distribution and content of selenium in soils. Acta Pedol. Sin. 42 (3), 410.

Hashemian, M., Hekmatdoost, A., Poustchi, H., Nasrabadi, F.M., Abnet, C.C., Malekzadeh, R., 2014. Systematic review of zinc biomarkers and esophageal cancer risk. Middle East J. Digest. Diseas. 6 (4), 177.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2. Springer.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10 (6), e0125814.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modelling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214–215, 141–154.

Humphrey, O.S., Young, S.D., Bailey, E.H., Crout, N.M.J., Ander, E.L., Watts, M.J., 2018. Iodine soil dynamics and methods of measurement: a review. Environ Sci Process Impacts 20 (2), 288–310.

Humphrey, O.S., Young, S.D., Bailey, E.H., Crout, N.M.J., Ander, E.L., Hamilton, E.M., Watts, M.J., 2019. Iodine uptake, storage and translocation mechanisms in spinach (*Spinacia oleracea* L.). Environ. Geochem. Health 41 (5), 2145–2156.

Humphrey, O.S., Young, S.D., Crout, N.M.J., Bailey, E.H., Ander, E.L., Watts, M.J., 2020. Short-term iodine dynamics in soil solution. Environ. Sci. Technol. 54 (3), 1443–1450.

Hurst, R., Siyame, E.W., Young, S.D., Chilimba, A.D., Joy, E.J., Black, C.R., Ander, E.L., Watts, M.J., Chilima, B., Gondwe, J., 2013. Soil-type influences human selenium status and underlies widespread selenium deficiency risks in Malawi. Sci. Rep. 3, 1425.

Johnson, C.C., 2003. Database of the iodine content of soils populated with data from published literature. In: British Geological Survey Commissioned Report CR/03/004N, p. 38.

Johnson, C.C., Ander, E.L., 2008. Urban geochemical mapping studies: How and Why we do them. Environ. Geochem. Health 30 (6), 511.

Joy, E.J., Ander, E.L., Young, S.D., Black, C.R., Watts, M.J., Chilimba, A.D., Chilima, B., Siyame, E.W., Kalimbira, A.A., Hurst, R., 2014. Dietary mineral supplies in Africa. Physiol. Plant. 151 (3), 208–229.

Joy, E.J.M., Broadley, M.R., Young, S.D., Black, C.R., Chilimba, A.D.C., Ander, E.L., Barlow, T.S., Watts, M.J., 2015. Soil type influences crop mineral composition in Malawi. Sci. Total Environ. 505, 587–595.

Kisinyo, P., Palapala, V.A., Gudu, S., Opala, P., Othieno, C., Okalebo, J., Otinga, A., 2014. Recent Advances Towards Understanding and Managing Kenyan Acid Soils for Improved Crop Production.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36 (11), 1–13.

Lagacherie, P., McBratney, A., Voltz, M., 2006. Digital Soil Mapping: An Introductory Perspective. Elsevier.

Lark, R., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51 (1), 137–157.

Li, X., Chang, S.X., Liu, J., Zheng, Z., Wang, X., 2017. Topography-soil relationships in a hilly evergreen broadleaf forest in subtropical China. J. Soils Sediments 17 (4), 1101–1115.

Liu, J., Wang, J., Leng, Y., Lv, C., 2013. Intake of fruit and vegetables and risk of esophageal squamous cell carcinoma: a meta-analysis of observational studies. Int. J. Cancer 133 (2), 473–485.

Matheron, G., 1962. Traité de géostatistique appliquée. Editions Technip.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

McGrath, S., Loveland, P.J., 1992. Soil Geochemical Atlas of England and Wales. Blackie Academic & Professional.

Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140 (4), 324–336.

Moebius-Clune, B., Van Es, H., Idowu, O., Schindelbeck, R., Kimetu, J., Ngoze, S., Lehmann, J., Kinyangi, J., 2011. Long-term soil quality degradation along a cultivation chronosequence in western Kenya. Agric. Ecosyst. Environ. 141 (1–2), 86–99.

Muindi, E., Mrema, J., Semu, E., Mtakwa, P., Gachene, C., Njogu, M., 2015. Phosphorus Adsorption and its Relation with Soil Properties in Acid Soils of Western Kenya.

Naimi, S., Ayoubi, S., Demattê, J.A.M., Zeraatpisheh, M., Amorim, M.T.A., Mello, F.A.d.O., 2022. Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning. Geocarto Intern. 37 (25), 8230–8253.

Neina, D., 2019. The role of soil pH in plant nutrition and soil remediation. Appl. Environ. Soil Sci. 2019, 5794869.

Noulas, C., Tziouvalekas, M., Karyotis, T., 2018. Zinc in soils, water and food crops. J. Trace Elem. Med. Biol. 49, 252–260.

Oliver, M.A., Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. Int. J. Geogr. Inf. Syst. 4 (3), 313–332.

Opala, P.A., Odendo, M., Muyekho, F.N., 2018. Effects of Lime and Fertiliser on Soil Properties and Maize Yields in Acid Soils of Western Kenya.

Otieno, H.M., Zingore, G.N.C.W.S., 2018. Effect of farmyard manure, lime and inorganic fertiliser applications on soil pH, nutrients uptake, growth and nodulation of soybean in acid soils of western Kenya.

Pebesma, E., 2018. sf: simple features for R. R J. 10 (1), 439–446.

Penn, C.J., Camberato, J.J., 2019. A critical review on soil chemical processes that control How soil pH affects phosphorus availability to plants. Agriculture 9 (6), 120.

Pisarek, P., Bueno, M., Thiry, Y., Nicolas, M., Gallard, H., Le Hécho, I., 2021. Selenium distribution in French forests: influence of environmental conditions. Sci. Total Environ. 774 (144), 962.

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92.

R Core Team, 2020. R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria.

Rawlins, B., 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.

Rayman, M.P., 2012. Selenium and human health. Lancet 379 (9822), 1256–1268.

Rickard, W.H., Price, K.R., 1984. Iodine in terrestrial wildlife on the U.S. department of energy's Hanford Site in southcentral Washington. Environ. Monit. Assess. 4 (4), 379–388.

Schaafsma, T., Wakefield, J., Hanisch, R., Bray, F., Schüz, J., Joy, E.J.M., Watts, M.J., McCormack, V., 2015. Africa's Esophageal cancer corridor: geographic variations in

incidence correlate with certain micronutrient deficiencies. PLoS One 10 (10), e0140107.
Seibert, J., Stendahl, J., Sørensen, R., 2007. Topographical influences on soil properties in boreal forests. Geoderma 141 (1–2), 139–148.
Sekulić, A., Kilibarda, M., Heuvelink, G., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens. 12 (10), 1687.
Shao, Y., Cai, C., Zhang, H., Fu, W., Zhong, X., Tang, S., 2018. Controlling factors of soil selenium distribution in a watershed in Se-enriched and longevity region of South China. Environ. Sci. Pollut. Res. 25 (20), 20048–20056.
Slessarev, E.W., Lin, Y., Bingham, N.L., Johnson, J.E., Dai, Y., Schimel, J.P., Chadwick, O. A., 2016. Water balance creates a threshold in soil pH at the global scale. Nature 540 (7634), 567–569.
Steevens, J., van den Brandt, P.A., Goldbohm, R.A., Schouten, L.J., 2010. Selenium status and the risk of esophageal and gastric cancer subtypes: the Netherlands cohort study. Gastroenterology 138 (5), 1704–1713.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modelling. J. Chem. Inf. Comput. Sci. 43 (6), 1947–1958.
Sylvain, J.-D., Anctil, F., Thiffault, É., 2021. Using bias correction and ensemble modelling for predictive mapping and related uncertainty: a case study in digital soil mapping. Geoderma 403 (115), 153.
Tan, J., Xie, X., Zuo, J., Xing, X., Liu, B., Xia, Q., Zhang, Y., 2021. Coupling random forest and inverse distance weighting to generate climate surfaces of precipitation and temperature with Multiple-Covariates. J. Hydrol. 598 (126), 270.
Towett, E.K., Shepherd, K.D., Tondoh, J.E., Winowiecki, L.A., Lulseged, T., Nyambura, M., Sila, A., Vågen, T.-G., Cadisch, G., 2015. Total elemental composition of soils in Sub-Saharan Africa and relationship with soil forming factors. Geoderma Region. 5, 157–168.
Vane, C.H., Kim, A.W., Beriro, D., Cave, M.R., Lopes Dos Santos, R.A., Ferreira, A.M., Collins, C., Lowe, S.R., Nathanail, C.P., Moss-Hayes, V., 2021. Persistent organic pollutants in urban soils of Central of London, England, UK: measurement and spatial modelling of Black Carbon (BC), Petroleum Hydrocarbons (TPH), Polycyclic Aromatic Hydrocarbons (PAH) and Polychlorinated Biphenyls (PCB). Adv. Environ. Eng. Res. 2 (2).
Wadoux, A.M.-C., Brus, D.J., Heuvelink, G.B., 2019. Sampling design optimization for soil mapping with random forest. Geoderma 355 (113), 913.
Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210 (103), 359.
Watts, M., Mitchell, C., 2009. A pilot study on iodine in soils of Greater Kabul and Nangarhar provinces of Afghanistan. Environ. Geochem. Health 31 (4), 503–509.
Watts, M.J., Middleton, D.R., Marriott, A.L., Humphrey, O.S., Hamilton, E.M., Gardner, A., Smith, M., McCormack, V.A., Menya, D., Munishi, M.O., 2019. Source apportionment of micronutrients in the diets of Kilimanjaro, Tanzania and Counties of Western Kenya. Sci. Rep. 9 (1), 1–14.
Watts, M.J., Middleton, D.R.S., Marriott, A., Humphrey, O.S., Hamilton, E., McCormack, V., Menya, D., Farebrother, J., Osano, O., 2020. Iodine status in western Kenya: a community-based cross-sectional survey of urinary and drinking water iodine concentrations. Environ. Geochem. Health 42 (4), 1141–1151.
Watts, M., Humphrey, O., Cave, M., Osano, O., Menya, D., 2021a. Western Kenya soil geochemistry. In: N.E.N.G.D. Centre (Ed.).
Watts, M.J., Menya, D., Humphrey, O.S., Middleton, D.S., Hamilton, E., Marriott, A., McCormack, V., Osano, O., 2021b. Human urinary biomonitoring in Western Kenya for micronutrients and potentially harmful elements. Int. J. Hyg. Environ. Health 238 (113), 854.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons.
Wragg, J., Cave, M., 2021. Modelling and mapping total and bioaccessible arsenic and lead in stoke-on-trent and their relationships with industry. Geosciences 11 (12), 515.
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), 1–17.
Xu, W., Zhu, J.-M., Johnson, T.M., Wang, X., Lin, Z.-Q., Tan, D., Qin, H., 2020. Selenium isotope fractionation during adsorption by Fe, Mn and Al oxides. Geochim. Cosmochim. Acta 272, 121–136.
Yu, M., Liu, Q., 2021. Deep learning-based downscaling of tropospheric nitrogen dioxide using ground-level and satellite observations. Sci. Total Environ. 773, 145145.
Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452.
Zhu, Q., Lin, H.S., 2010. Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. Pedosphere 20 (5), 594–606.