

**CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE**



**Czech University  
of Life Sciences Prague**

**Faculty of Agrobiography, Food and Natural Resources**

**Department of Soil Science and Soil Protection**

Digital soil mapping methods of soils properties and soil nutrients

Doctoral dissertation

Author: **MSc. Kingsley John**

Supervisor: **doc. Ing. Vít Penížek, Ph.D.**

Consultant: **RNDr. Tereza Zádorová, Ph.D.**

**Praha 2 0 2 2**

## **ACKNOWLEDGEMENT**

I am delighted to express my profound gratitude to doc. Ing. Vít Penížek, Ph.D and RNDr. Tereza Zadorova, Ph.D., for their supervision and general guidance of the thesis. I would like to thank Prof. Dr. Ing. Luboš Borůvka for the opportunity to learn under his quiver of knowledge. I want to appreciate the opponents, prof. Dr. Ing. Bořivoj Šarapatka, CSc., Mgr. Jan Skála, Ph.D and Ing. Lukáš Brodský, Ph.D for their immense comments on the first version of my thesis; I am sincerely humbled after my first doctoral defence attempt. I also wish to express gratitude to the co-authors of the featured publications for their contributions and to all other members of staff at the Department of Soil Science and Soil Protection who gave their assistance. Finally, I wish to thank my wife, Dr. Esther John Kingsley, my siblings, and my parents for their support throughout my academic studies. And above all, I want to thank the Almighty God for the strength, wisdom, and motivation throughout my academic journey.

## PREFACE

The selected publications presented in this thesis were compiled as part of my research activities between 2018 and 2022. All the studies are interconnected to the subject matter, digital soil mapping of soil properties and nutrients. The studies were conducted in Nigeria, the Czech Republic, and Morocco, characterized by different climatic conditions, land use, and geological forms.

All the research was performed at the Department of Soil Science and Soil Protection at the Czech University of Life Sciences in Prague in cooperation with the Department of Soil Science, University of Calabar, Nigeria, National Institute of Agricultural Research (INRA), Morocco, and, as part of various grant obligations, independent research, or part of the thesis work. Grant providers and co-authors are acknowledged within the respective publications. During the doctoral study, a central team evolved. Members were responsible for significant contributions to the conceptualization, consultation, and development of the research and the practical aspects such as sampling and analysis. Consistent primary contributors to the listed research papers are Prof. Chengzhi Qin, Ndiye Michael Kebonye, Ph.D., Ing. Prince Chapman Agyeman, and Isong Abraham Isong, Ph.D., with a project overview and thesis supervision by doc. Ing. Vít Penížek and RNDr. Tereza Zadorova, Ph.D.

Digital soil mapping introduction into Soil Science is still significantly evolving from the research phase into the global development and creation of soil maps. Approaches and methods to improve digital soil mapping of soil properties to ensure the accurate representation of the reality of the soil situation still border around financial budget, optimum sample size, available soil data, and expert knowledge. Therefore, this current research work reveals pragmatic models in predicting soil properties with few samples and ways to improve modelling by simulating sample ratios with different predictive models to cover individual needs where necessary. The present doctoral thesis applied a few sample ratios from existing databases to model some selected soil nutrients in low relief using a more pragmatic kriging model (John et al., 2020). We compared kriging and inverse weighting distance interpolations (John et al., 2020) and tested the combination of cokriging and Gaussian process regression in modelling soil sulphur (John et al., 2021). We also attempted different multiple linear model functions in estimating soil organic matter via some selected soil physical properties (Ofem et al., 2020). Furthermore, we evaluated the sample ratio and sampling schemes for predicting soil nutrient elements (John et al., 2022).

The research initiation and development approach considers the workability of digital soil mapping models with few samples, improvement of machine learning model, variability of sampling strategies and sample ratios, and the influence of different linear functions on soil modelling. And while focusing on these challenges allowed for the synergizing of a compilation of holistic study during the Ph.D research.

**John, K.**, S. M. Afu, I. A. Isong, E. E. Aki, N. M. Kebonye, E. O. Ayito, P. A. Chapman, M. O. Eyong, and V. Penížek. "Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics." *International Journal of Environmental Science and Technology* (2021): 1-16.

**John, K.**, Afu, S. M., Isong, I. A., Chapman, P. A., Kebonye, N. M., & Ayito, E. O. "Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the Southeastern soils of Nigeria." *Environmental Engineering & Management Journal (EEMJ)*, 20(7).

**John, K.**, Agyeman, P. C., Kebonye, N. M., Isong, I. A., Ayito, E. O., Ofem, K. I., & Qin, C. Z. "Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur." *CATENA*, 206, 105534.

Ofem, K.I., **John, K.**, Pawlett, M., Eyong, M.O., Awaogu, C.E., Umeugokwe, P., Ambrose-Igho, G., Ezeaku, P.I. and Asadu, C.L.A., 2021. Estimating Soil Organic Matter: A Case Study of Soil Physical Properties for Environment-Related Issues in Southeast Nigeria. *Earth Systems and Environment*, 5(4), 899-908.

**John, K.**, Bouslihim, Y., Bouasria, A., Razouk, R., Hssaini, L., Isong, I. A., Ayito, E. O., Ambrose-Igho, G. (2022). Assessing the impact of sampling strategy in Random Forest-based predicting of soil nutrients: a study case from Northern Morocco Geocarto International (2022): 1-14.

## TABLE OF CONTENTS

1.	Literature review	1
1.1	Background of the study	1
1.2	Conventional mapping technique	3
1.3	Digital soil mapping (DSM)	4
1.3.1	Digital soil mapping approaches	4
1.3.2	Predictive models in DSM	5
1.3.3	Geostatistics and deterministic interpolation	6
1.3.4	Machine learning (ML) models	8
1.4	Environmental covariates	12
1.5	Soil nutrient and soil properties variabilities	14
1.6	Sampling strategies and sampling ratio	15
2.	Areas of study	19
2.1	Cross River State, Nigeria	21
2.2	Frydek-Mistek district, Czech Republic	23
2.3	Taounate province, Morocco	24
3.	Aims and hypotheses	26
3.1	Aims	26
3.2	Hypotheses	27
4.	Synthesis and concluding remarks	28
4.1	Synthesis of key findings	28
4.2	Concluding remarks	30
	References	32
	Publication list	47

# **1. LITERATURE REVIEW**

## **1.1 Background of the study**

Soil nutrients are vital for soil fertility and the development of crops (Lal, 2015). Therefore, investigating, modelling, and mapping the spatial distribution of soil nutrients is essential for practical farming and sustainable land management (Ma et al., 2017). And to further improve soil nutrient levels, soil management with an appropriate understanding of soil properties variability is required. Soil properties are the characteristics of a given soil and are considered crucial in the availability and mobility of soil nutrients (Bardgett and Wardle, 2010). Besides that, farming management techniques, such as irrigation and fertilization, and soil formation variables, such as soil parent materials, impact soil properties and nutrient spatial variability (Davatgar et al., 2012). As a result, managing agricultural areas as a single unit may cause soil deterioration by treating regions with high nutrient content with an excess of inputs and those with low nutrient content with insufficient input materials (Ferguson et al., 2002).

In some parts of the world and sub-Saharan Africa, there is a shortage of understanding of the spatial variability of different soil fertility conditions, such as soil acidity and nutrient deficits, which is a crucial impediment to establishing appropriate liming and fertilizer recommendations. Furthermore, the variability of soil nutrients is a significant constraint for sustainable crop production due to the resulting non-uniformity of output across different field sections. Soil nutrient is more nature-driven compared to, e.g. western Europe, where intensive agriculture is carried out, so it may also vary due to natural processes and soil management types. Spatial variation in soil nutrients on crop yields is evident in cultivated sloping fields. However, there is currently a paucity of information about the spatial variation of soil nutrients (Ge et al., 2007). By utilizing appropriate soil management techniques, it will be feasible to reduce the detrimental impacts of soil nutrient spatial variation on agricultural productivity.

Soil mapping has traditionally been based on time-consuming data collection, field surveys, interpretation, field verification, demarcation, and mapping (Scull et al., 2003). The fast change in climate, land surface and eco-hydrological modelling in the last 20 years necessitates maps of soil attributes with high resolution and low uncertainty. This requirement can no longer be met by traditional soil mapping. Soil scientists worldwide have employed digital soil mapping (DSM) approaches to tackle this problem by constructing statistical models based on soil measurements, environmental factors, and statistical models (Lagacherie and McBratney, 2006; Minasny and

McBratney, 2016). And this approach has been employed to map soil nutrients spatially and properties within a field, relying upon using soil environmental covariates such as soil properties (from existing soil information), remote sensing data, digital elevation model (DEM), micro-climatic data, and geology (Zeraatpisheh et al., 2020; Mosleh et al., 2016).

Digital mapping of soil attributes is required to inform soil and land use management. Since the idea of digital soil mapping (McBratney et al., 2003), many studies have been conducted to increase and deepen our understanding of accurate soil spatial prediction (Minasny and McBratney, 2016). This is because digital soil mapping will be demand-driven rather than supply-driven for land management applications, with operational uses of digital soil maps for land use planning (Kidd et al., 2020; Searle et al., 2021). However, local and regional knowledge and assessment of soil maps are required for both theoretical and practical uses (Arrouays et al., 2017; Pásztor et al., 2020).

When generating maps of soil properties and nutrients for specific land management issues, one of the associated challenges is the high number of sampling points necessary to produce accurate maps of soil physical and chemical properties, which intensely adds to the expense of the mapping process. One of the methods to solve this problem is by developing methods that require a lower number of sampling points to produce accurate maps. This new approach should be flexible enough to use inexpensive covariates sampled at high density, for example, crop yield, vegetation index, apparent soil electrical conductivity, etc. Machine learning (ML) has recently proved to be an efficient technique for predicting and mapping soil attributes (Khaledian and Miller, 2020; Nabiollahi et al., 2021; Nyéki et al., 2021). ML can be understood as the automated process of learning by algorithms based on large datasets. These algorithms are efficient for working with large volumes of data. Thus, an ML model can be used in data mining, pattern recognition, regression, and classification (Heung et al., 2016; Khaledian and Miller, 2020; Liakos et al., 2018; Parmley et al., 2019). Several ML studies have been employed in predicting and spatial distribution analysis of soil properties (Chen et al., 2019; Nyéki et al., 2021; Shaddad et al., 2016). Most of these ML studies involved mapping large areas with high attribute variations. On the other hand, hybrid methods have also been introduced, which include the combination of geostatistical techniques and machine learning (e.g., linear regression residual kriging, linear regression residual inverse distance weighting), and the interpolation accuracy can be further improved by correcting the residuals in the global interpolation method). Therefore, this thesis aims to investigate whether

geostatistic techniques, ML, or the combination of the two approaches are relevant and useful in mapping soil properties and nutrients at the field scale used with varied sample ratios.

## **1.2 Conventional mapping technique**

The conventional mapping technique is one of the significant sources of soil spatial information via topographical details. Traditional soil maps are generally produced using a free survey (Kempen et al., 2012). In this survey, the soil surveyor uses a soil-landscape model to select suitable observation locations via a landscape equation or concept (Bregt, 1992; Hudson, 1992). In obtaining soil information from a given area, the soil mapper first delineates the site through ground-truthing to establish a soil-landscape model (Zhu, 2000). The soil-landscape model encapsulates the relationship between soils in the location and the different land positions or units. Next, the soil surveyor manually sketches the map spatial extents of different soils or combinations of soils through photo-image analysis. And the output results of the soil units are then represented using polygons. The individual areas on the maps are then referred to as map units (Lark and Beckett, 1998) and each formed with a polygon depicts the spatial arrangement thereof. It is an effort to map units as much as possible to match one classification unit, which is then used in the map legend. When a given land is included in a map or classification unit, it is said that it is a typical representative. That's why the polygonal approach often limits an accurate description of soil cover (Zhu, 2000) and reduces the possibility of capturing continuous changes in soil properties. The polygon-based mapping practice is based on the discrete conceptual model (Zhu, 1997), limiting the soil mapper's ability to produce accurate soil maps. Traditional soil maps are all-purpose maps: they yield information on the three-dimensional spatial distribution of a wide range of soil properties that are interpreted from representative soil profile descriptions associated with the map units.

The problems associated with conventional soil mapping are that the size of the soil body can be represented as a polygon is limited on paper (Kempen et al., 2012). The polygons represent only the distribution of a set of prescribed soil classes and limit the ability to update soil surveys rapidly and accurately. In addition, the conventional process involves detecting different soil formation processes, i.e., information on the map unit composition, soil profile descriptions, and map unit interpretations (Soil Survey Staff 2014; Beaudette and O'Geen, 2009).



## 1.3 Digital soil mapping

### 1.3.1 Digital soil mapping approaches

Digital soil mapping involves generating and populating soil information systems via mathematical models to infer soil types and properties' spatial and temporal variability from observed soil data and knowledge developed from environmental covariates (Lagacherie and McBratney, 2007). The DSM technique is centred on the SCORPAN model proposed by Mcbratney et al. (2003). The model allows incorporating soil information as a covariate via the soil-forming factors represented by S (McBratney et al., 2003). SCORPAN approach involves incorporating a limited number of field measurement data to a large geographical area and estimating the targeted soil properties over the whole study area. SCORPAN model is a modification of Jenny's soil-forming equation (Jenny, 1941),

$$S = f( cl, o, r, p, t) \quad \text{Equation (1)}$$

where S = soil, cl = climate, o = organisms (including humans), r = relief, p = parent material and t = time of formation. Jenny's equation was modified to describe the impact of the environment on soil formation and development and quantify the process by applying the mathematical concepts through digital mapping (McBratney et al., 2003). The SCORPAN model, which captures the current approaches of DSM, is described as follows,

$$Sc, p = f( s, c, o, r, p, a, n) + e \quad \text{Equation (2)}$$

where s = soil, other properties or prior knowledge of the soil at a point; c = climate, climatic properties of the environment at a point; o = organisms, vegetation or fauna or human activity; r = topography, landscape attributes; p = parent material, lithology; a = age, the time factor; n: space, relative spatial position; e = autocorrelated random spatial variation, Sc = is the soil classification unit (e.g., soil type), and Sp is the soil property. The land value represents, for example, from previous work or remote sensing data. Other factors are represented by continuous variables, such as the average annual temperature, rainfall, etc. Henderson et al. (2005) exemplified this by using existing legacy soil class mapping for predicting many soil properties across the Australian continent without kriging the residuals. McBratney et al. (2003) outlined that the success of the SCORPAN models is dependent on (a) a sufficient number of additional data (in terms of the number of variables and the number of sampled points), (b) a sufficient amount of data on the soil,

(c) the existence of a function that can describe the relationship of the soil and the additional data, and (d) a good correlation between the soil (or its properties) and the environment.

Digital soil mapping models include a wide range of methods, such as geostatistical models (Lark et al., 2006; Goovaerts, 2011), tree models (Bui et al., 2006; Connolly et al., 2007), neural networks (Behrens et al., 2010), fuzzy systems (Zhu et al., 2000; Odgers et al., 2011; Yang et al., 2013) and ensemble machine learning models (Hengl et al., 2021; Sylvain et al., 2021; Brungard et al., 2021). Furthermore, most DSM applications are research-oriented and have regional specifics (Kempen et al., 2010; Kempen et al., 2012). As a result, the active engagement of DSM is still limited (MacMillan et al., 2007; Lilburne et al., 2012; Grunwald, 2009; Grunwald et al., 2011). The challenges of DSM techniques border around larger spatial extents, data availability of important soil information and large datasets. For example, if the sample size and the number of prediction locations are wide, geostatistical models are computationally intensive (Cressie and Johannesson, 2008). Also, in geostatistics, modelling the non-linear relationship between a soil attribute and many cross-correlated variables is complex and presents new obstacles (e.g. many parameters have to be estimated). However, Lark et al. (2006) showed that the combined fixed and random effects by residual maximum likelihood estimation deal with the non-stationary variance. At the same time, the machine learning models can establish a non-linear relationship between soil property and auxiliary attributes. The model requires high-resolution environmental covariates and readily available environmental covariates for a good prediction of a targeted soil property. However, in the case of machine learning models, it is worth noting that developing a specific data infrastructure is recommended to obtain a robust prediction output. Besides that, DSM studies have shown the potential to provide better and more accurate information on the spatial variability of a targeted soil (Scull et al., 2003)

### 1.3.2 Predictive models in DSM

The various predictive models adopted in this project were used to meet the specific needs of the area under investigation. For example, little is known about predictive models employed in digital soil mapping in Nigeria, especially in the southeastern part. Therefore, we attempted to test widely used interpolation models with corresponding soil points available to demonstrate the importance of these tools in accurate soil studies, which can also be used for educational purposes. In the Czech Republic, the model applied for the study was to add to the existing work of literature on digital soil mapping by introducing hybrid machine learning models. While for Morocco, we

attempted to establish a more robust ML approach to provide an optimum sample ratio to avoid unnecessarily wasting resources as soil databases grow in the region.

Some of the following models have been employed in developing DSM techniques in serving site-specifics, regional, national, and global issues and were used in our study:

### 1.3.3 Geostatistics and deterministic interpolation

The geostatistical methods, which essentially involve the kriging methods, are the earliest DSM methods applied to model soil properties (Odeh et al., 1995; Gessler et al., 1995; McKenzie and Ryan, 1999). One of the drawbacks of kriging in soil mapping is that it ignores environmental variables that are known or expected to be connected with target soil variables, such as Jenny's conceptual equation's soil-forming components or the modified model (i.e., SCORPAN). This approach involves the quantitative description of the spatial variation of soils. However, the means of incorporating the knowledge from the soil-forming model to accurately estimate soil properties was lacking. The challenge resulted in the development of universal kriging (Matheron, 1969), which has a significant problem; one needed the variogram of the random residuals from the drift, and one could not obtain those residuals without knowing the variogram. Nonetheless, a rough-and-ready empirical approach appeared to be fitting the target variable on values of the covariates using ordinary least-squares regression, geostatistically analyzing the residuals, and then adding back the regression predictions to the kriged prediction of the residuals. The regression kriging model has been extended in many ways, including three-dimensional and space-time mapping (Gasch et al., 2015; Cappello et al., 2021) and Bayesian and generalized linear modelling (Steinbuch et al., 2022). Geostatistics has proven effective for measuring spatial variability of soil characteristics, and soil scientists and agricultural engineers have increasingly used it (Webster and Oliver, 2001). Semivariograms and cross-semivariograms have been used to define and model spatial variation of data to analyze how data points are connected to separation distances. In contrast, kriging employs modelled variance to estimate values across samples (Journel and Huijbregts, 1978). The application of the geostatistical interpolation method help reduces the costs of field sampling and laboratory analysis, given that a set of soil observation points sufficiently represents the study area. There are different geostatistical methods such as kriging (e.g., ordinary kriging, simple kriging, universal kriging, cokriging, empirical Bayesian kriging, and others) and hybrid kriging model (regression kriging). Kriging interpolation yields the best linear unbiased estimates and information on the estimation error distribution and shows solid statistical

advantages. Cokriging (Cok), simple kriging (SK), and regression kriging (RK) are a few of the advanced and hybrid geostatistical tools which consider the relationship between primary and secondary variables. The difference between SK and OK is that the global mean in OK is unknown, while it is known in SK (however, it is unrealistic). As a result, when residuals have been computed, the known mean ( $m$ ) is added back to the data using SK. Estimating the values as a departure from the global mean is more suitable since we know a random variable's deterministic component. Other significant distinctions between SK and OK are based on the global mean assumptions of known and unknown. Cok is an extension of kriging that may be used when two or more variables are spatially linked. The following generalization uses just one co-variable for ease of understanding (e.g. Heisel et al., 1999). The primary variable (the variable of immediate interest, such as soil qualities) and the co-variable are weighted and averaged in Cok. Odeh et al. (1995) claimed that in the RK type, the regression residuals indicate uncertainty and are considered by the kriging systems. And to get at a target variable, a regression must be run first, and only then may kriging be applied with the injection of regression errors as prediction uncertainty. The idea is that kriging after regression may enhance prediction performance in contrast to when regression or kriging are done separately (by introducing the uncertainty due to regression mistakes into kriging equations). For example, Bangroo et al. (2020) applied RK and OK to predict SOC and total nitrogen (TN), revealing that RK has better prediction accuracy. Some scientific works have also shown the efficacy of Cok over other methods and found Cok outperforms other methods due to the inclusion of environmental covariates (Hooshmand et al., 2011; Singh et al., 2016; Tziachris et al., 2017). Besides, the novel empirical Bayesian kriging (EBK) has proven more robust and pragmatic in making an accurate prediction. Empirical Bayesian kriging (EBK) is an advanced geostatistical prediction technique that combines kriging and linear mixed model to evaluate precise models at a local scale (Schabenberger and Gotway, 2017). EBK differs from other kriging interpolations because it considers the uncertainties related to variogram plotting. Also, it automates the most challenging aspects of composing an adequate kriging model (Krivoruchko and Gribov, 2019). It can represent the stochastic spatial process locally as a stationary or non-stationary random field, where the parameters of the locally defined random field vary across space. In Bayesian Kriging, Bayes' theorem is adopted to integrate prior knowledge to produce posterior distribution taking into account the uncertainty in covariance function parameters. EBK is efficient in other fields of land and terrestrial science in spatial prediction (Giustini et al., 2019; Li et al., 2020; Lima et al., 2021). EBK does not require specification of the prior distribution for

model parameters; it allows moderate local and large global data non-stationarity, locally transforms data to Gaussian distribution if needed, works reasonably fast, and produces reliable outputs with default parameters (Krivoruchko and Gribov, 2019). In addition, this geostatistical model potentially outperforms other classical geostatistical models and is advantageous, particularly when common geostatistical modelling assumptions are violated (Gribov and Krivoruchko, 2020; Pilz and Spöck, 2008). Mirzaei and Sakizadeh (2015) reported that the EBK model performed better in predicting groundwater contamination than OK and IDW, respectively. Also, Hussain et al. (2014) opined that EBK is the most suitable geostatistical method for spatial prediction of total dissolved solids in drinking water.

Another interpolation method incorporated in the DSM approach is the deterministic model (e.g., inverse distance weighting). This interpolation method creates surfaces from observation points based on the extent of similarity (inverse distance weighting) or the degree of smoothing (Radial Basis Functions). A value for an attribute at an unsampled location is assumed to be a linearly weighted average of known data points occurring in the unsampled location's general neighbourhood. It is an accurate interpolator, and it is said that the sample density should be high in comparison to the local variance in the data for it to produce the best results (Burrough and McDonnell, 1998). Also, this is one of the easiest interpolation methods to be applied and has been shown to generate good predictions for unsampled values (Wong, 2017). For example, Gotway et al. (1996) obtained better results with inverse distance weighting (IDW) to kriging predicting soil organic matter and nitrogen contents. Also, IDW performed slightly better than the OK model in the spatial prediction of pH and soil organic matter in Western Australia (Robinson & Metternicht 2003).

#### 1.3.4 Machine learning (ML) models

According to Hartemink et al. (2008), machine learning (ML) approaches are a broad category of non-linear data-driven algorithms initially used for data mining and pattern recognition. Still, they are now widely applied to regression and classification applications across many scientific disciplines. Unlike geostatistical approaches, where alteration of the original observations is frequently necessary to meet the assumptions, ML algorithms do not assume the distribution of the observations. ML models can also handle large numbers of cross-correlated variables as predictors.

ML incorporated into the DSM approach could either be supervised or unsupervised. According to Sathya and Abraham (2013), supervised machine learning is based on training a data sample from a data source that has previously been classified correctly. In contrast, unsupervised machine learning can learn and organize information without receiving an error signal to evaluate a potential solution. In DSM, most models are centred on supervised learning (Russell and Norvig, 1995): understanding the associations between the targeted soil property and independent variables based on training samples and their environmental covariates. Large training samples are always required for supervised ML models. However, the labour-intensive field sampling campaign often limits the number of samples (Zhang et al., 2021). Inadequate sample data may limit the ability of supervised learning algorithms to learn. Supervised machine learning has been widely applied in DSM, and these models include random forest (RF) (Breiman, 2001; Heung et al., 2014; Hengl et al., 2015), multiple linear regression (MLR) (Forkuor et al., 2017; Chen et al., 2020), cubist (Quinlan, 1992), Gaussian process regression (GPR) (Xue et al., 2020), support vector machine (e.g. Were et al., 2015), and artificial neural networks (e.g. Behrens et al., 2005). These ML models have been used in mapping specific soil properties (Giasson et al., 2015; Minansny and Mcbratney, 2006; Bui and Moran, 2003; Henderson et al., 2005; Bui et al., 2021). The vastly applied ML in DSM are cubist and random forest (Breiman, 2001; Quinlan, 1992). This is because both models approach subsets data by rules related to the predictor variables and fit a linear regression model to each subset (Appelhans et al., 2015). Also, RF, as well as cubist, can be easily interpreted based on the relative importance of the modelling procedure (Walton, 2008). Both models have been successfully applied to map continuous and categorical soil properties on both a regional and national scale of map (Grimm et al., 2008; Guo et al., 2015; Rossel et al., 2015; Mulder et al., 2016; Liang et al., 2019).

The cubist model is developed as an extension of the M5 tree model (Quinlan, 1992). According to Kuhn (2008), the model structure consists of a conditional component—or piecewise function—acting as a decision tree coupled with multiple linear regression models. The Cubist method’s main benefit is to add multiple training committees and boosting to make the weights more balanced. The Cubist model adds boosting with training committees (usually greater than one) which is similar to the method of “boosting” by sequentially developing a series of trees with adjusted weights. The number of neighbours in the Cubist model is applied to amend the rule-based prediction (Kuhn, 2008). The “cubist” function in the CARET package in R software can be implemented to perform the model (R Core Team, 2019).

Random forests (RF) is a decision tree ensemble classifier built on regression trees with random inputs that partition sub-datasets and predict variables (Breiman, 2001). It can handle both continuous and categorical variables. According to Heung et al. (2014), RF is accurate as or better than adaptive boosting yet computationally faster. Also, the RF algorithm is robust to noise in predictors and thus does not require a pre-selection of variables (Díaz-Uriarte et al., 2006). In RF, the input response variables are randomly split into many small datasets in order to grow trees, and the input explained variables are randomly divided into each small dataset. No tree in RF uses the entire dataset and all predictive variables to fit each regression tree so that the tree can grow as deep as possible without pruning. The function 'randomForest' in the R package 'randomForest' can fit either a classification or a regression tree using Breiman's technique (Breiman, 2001). Three variables can determine the goodness of the model fitting: mtry (splits number), nz (node size), and ntree (tree number). 'mtry'. Different combinations of these three items can be used to find the best model.

Multiple linear regression (MLR) models aim to explain the spatial distribution of a dependent variable (e.g. targeted soil property) through a linear combination of predictors (independent variables such as environmental covariates) (Forkuor et al., 2017). The R basic package 'stats' offers a function 'lm' to fit linear models. Also, MLR can be performed via a cross-validation function in R (e.g., leaps and stepAIC functions) available in R's leaps and MASS packages. The leaps package in R is composed of "leapBackward", which fits a linear regression with backward selection, and "leapForward", with fittings for linear regression with forward selection. The "leapSeq" fits a linear regression with stepwise selection, while in stepAIC (also referred to as direction) (James et al. 2014). The leaps package is new in R and has been rarely applied in soil studies. The most frequent method for soil data modelling is the stepAIC linear regression model. On the other hand, the link between soil and auxiliary factors isn't always linear, and it's often unknowable and noisy (Hengl et al., 2004). Also, the challenge in applying regression models is the problem of multicollinearity, which happens when there is a significant correlation between the predictors (environmental covariates) (Forkuor et al., 2017).

Gaussian process regression (GPR) is a powerful tool in digital soil mapping when multiple explanatory variables are available (Xue et al., 2020). The model can be used for classification, regression, and fit models and requires no tuning parameter (Kuhn, 2008). Compared to other ML models, GPR yields well-defined confidence intervals, which are very important for soil scientists to evaluate the robustness of a model. GPR allows spatial interpolation and ancillary features to

create a model (Xue et al., 2020). Gaussian process regression (GPR) can be an efficient model for DSM mapping continuous soil properties. However, there is still a research gap in exploiting the model potentials, especially in contrast to geostatistical models and assessing their uncertainties. The GPR model, a novel ML model, can handle uncertainties in given measurement and unevenly spaced and correlated training samples through a user-specified covariance kernel. Above all, the main challenge in ML techniques is that the predicted value at each point is derived based on the predictor variables without considering the spatial autocorrelation of the data (Takata et al., 2007). Also, when there are few covariates, available environmental variables are only weakly correlated to the target variable, and insufficient data for calibration, ML models confront a hurdle. It's also important to remember that training a model solely on a regression matrix of paired observations of dependent and independent variables while neglecting their spatial interrelations is inherently suboptimal. Hence, it is essential to note that samples with spatial intelligence of a given field may improve the prediction accuracy of a targeted soil property or soil nutrients. Metcalfe et al. (2016) demonstrated how scale-dependent relations between weed density and soil properties could be examined with appropriate sampling and analysis. This may be achieved through an optimal sampling scheme and sampling ratios showing spatial variability in a given field. Because soil spatial variation is the product of multiple processes functioning at multiple geographical scales, changes in some soil properties can be patchy. Besides that, hybrid predictive models could be adopted for accurate prediction. However, these models do seem not data specific, but their models are. Therefore, the best model should be developed according to an individual situation based on the output obtained from the different models. It is possible to create hybrid models by combining different predictive models to improve the modelling performances. Further details on the framework for the other rules for the combination of classifiers were reported by Kittler et al. (1998).



## 1.4 Environmental covariates

Machine learning models are used to create digital representations of spatial soil distribution utilizing point soil measurements and spatially comprehensive environmental covariates in digital soil mapping (McBratney et al., 2003, Scull et al., 2003, Florinsky, 1998). Soil point measurements are response variables, and environmental covariates are predictors; both are available and can be employed in the digital soil mapping approach. Environmental covariates include terrain properties (obtained from digital elevation models) (Mueller and Pierce, 2003), remote sensing imagery (Wu et al., 2009), climatic data (Mishra et al., 2010) and soil data (Nussbaum et al., 2014; John et al., 2020). The digital elevation model (DEM) provides many data attributes to help soil scientists map and quantify landforms and soil variability (Wilson and Gallant, 2000). The source of the elevation data includes the techniques for measuring elevation either on the ground or remotely, the locations of samples and the density of samples, and the algorithms used to calculate different terrain attributes (Theobald, 1989; Chang and Tsai, 1991; Bolstad and Stowe, 1994; Florinsky, 1998; McKenzie et al., 2000). In addition, other remotely sensed data contain extractable soil information, e.g., spectral reflectance. These data produce reliable spatial-temporal information and offer possibilities of supplementing or reducing conventional soil sampling in soil mapping (Forkuor et al., 2017; Forkuor et al., 2017; Malone et al., 2016). RS data are readily available and are free of charge (e.g., Landsat, SRTM, Sentinel-1, -2) (Mulder et al., 2011).

Besides that, other secondary/auxiliary data can be found. For example, portable X-ray fluorescence spectroscopy (pXRF) and Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) have been employed to predict different response variables. For instance, Kebonye et al. (2021) used data sourced from portable X-ray fluorescence spectroscopy (pXRF) to map arsenic via regularized linear models. Their findings showed pXRF as a promising tool for estimating arsenic in the floodplain area of Příbram (Czech Republic). Also, John et al. (2021) revealed that the pXRF dataset could be promising in estimating soil organic carbon in the floodplain area of Příbram. At the same time, Agyeman et al. (2022) estimated nickel concentration via the Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) dataset using empirical Bayesian kriging and support vector machine regression models.

Nevertheless, there are significant flaws in applying the machine learning model to estimate soil spatial variability via different environmental covariates. First, adequate and evenly dispersed point soil data throughout the mapped region is required (Carré et al., 2007). Second, unlike soil-

landscape process models, the model structure looks at the empirical link between environmental covariates and soil parameters (Grunwald, 2009). Lastly, the variables are rough estimates of the natural environmental condition that shaped the soil. And where these challenges exist, soil researchers use the available observation points with readily accessible environmental covariates combined with robust predictive models in creating or updating soil property maps for different uses.

In general, mapping soil properties via environmental correlation involves the iterative development of predictive models for the location under investigation. Geology and pedologic soil formation may also serve as environmental covariates as they help explain the spatial variability of soil properties within or among agricultural fields. Besides this, soil variability may also be generated by tillage, and soil management activities may serve as factors employed in a DSM model to explain the variability of a targeted soil property. Nevertheless, these variables interact at different geographical and temporal scales, being locally influenced by erosion and deposition processes (Iqbal et al., 2005). The extraordinary complexity and range of spatial and temporal scales over which soil-forming processes operate make developing models for quantitative, mechanical, and mathematical spatial prediction an almost impossible task in routine soil surveys. However, there have been some notable attempts (Dietrich et al., 1995). Despite the complexity, a simplifying hypothesis is necessary, and reliance must be placed on approximate local models of pedogenesis with varying levels of empiricism. McSweeney et al. (1994) reported that soil-terrain modelling techniques had been developed as a quantitative method for predicting soil variability using observed patterns in environmental variables known to influence soil property variabilities, such as topography, hydrology, or geology. DSM techniques have been used to model the spatial distribution of specific soil properties, including A-horizon thickness, organic matter content, extractable P, pH, and sand and silt content (John et al., 2021; John et al., 2020; Moore et al., 1993), A-horizon thickness and depth to carbonates (Bell et al., 1994) and A-horizon thickness and solum depth (Gessler et al., 1995). Bell et al. (1992) predicted and mapped soil drainage class using topographic information derived from DEM, a perennial stream, ephemeral surface drainage paths, and geology. Penížek and Borůvka (2006) examined the influence of terrain derivatives on soil depth via cokriging and regression kriging methods. Their study reported that slope, aspect, and elevation influenced soil depth distribution. Similarly, Penížek et al. (2016) examined the influence of different terrain model resolutions on colluvial soils.

## 1.5 Soil nutrient and soil properties variabilities

In space and time, soil nutrients and properties display a high degree of variability (Fanuel et al., 2018). Because soil properties and nutrients result from the simultaneous interplay of biological, chemical, and physical processes working at many scales, such variability is continuous across the landscape and scale-dependent (Haileselassie et al., 2011; Panday et al., 2018). In most tropical, Mediterranean and temperate intensive farming systems, it is increasingly concerned that if the nutrient and soil organic matter (SOM) supplies continue to decline, the land's ability to support agriculture will be jeopardized (Lal, 2015). As a result, it is vital to comprehend the controllable physical characteristics and the related soil nutrient cycle mechanisms. For example, recent studies have found falling trends in primary productivity and SOM in Malawi owing to continuous use (Li et al., 2017; Messina et al., 2017; Mpeketula, 2016), as well as low nitrogen and soil organic carbon levels in Nigeria (Nafiu et al., 2012; John et al., 2019). On the other hand, these and other studies serve as pointers since they were done at national or point sizes and are not typical of soil conditions across and within farming landscapes (Forkuor et al., 2017).

Soil nutrient elements such as total soil nitrogen, available phosphorus, and exchangeable bases in western Africa, the central European region, and north-western Africa have been well-studied (John et al., 2020; Žížala et al., 2021; Al Masmoudi et al., 2021). In the western African region, for example, Nigeria, soil properties such as particle sizes, organic carbon, and pH are regularly studied via the conventional soil mapping approach (Esu, 2005). The north-western, for example, Morocco region, is currently developing its soil fertility map via DSM. On the other hand, the DSM approach is already established in the central European area (e.g., the Czech Republic) owing to its large-scale soil legacy data (Zádorová et al., 2020).

The spatial variability of soil properties and nutrients across different areas is influenced by environmental factors such as soil type, climate, vegetation, etc. For example, in Nigeria, according to Esu (2005), total nitrogen is generally low, ranging from 0.03 to 0.24 %. In southern Nigeria, which is predominantly rainforest, the total nitrogen content in soil is higher than in the Northern part of Nigeria (i.e., savannah). Conversely, due to the vegetation variation, available phosphorus is higher in the savannah region (i.e., northern Nigeria) than in the rainforest area (southern Nigeria). The available phosphorus ranges from 3 to 20 mg/kg. The current situation of the country current situation is due to natural ecosystem degradation following deforestation, overgrazing, nutrient mining, soil erosion, loss of biodiversity, and lack of up-to-date soil property map for accurate soil management decisions.

According to the National Institute of Agricultural Research (INRA), in Morocco, soil organic matter (SOM) ranges from 1.5 to 3.5 % in all the regions. The predominant SOM value ranged from 1.5 to 2.5%. The part with the dominant high SOM (2.5 to 3.5 %) is the Taounante, characterized by a temperate climate and stable slope. The high SOM values correspond to the slightly acid-to-alkaline reaction (5.5 to 8.5) and relatively high soil available phosphorus (10 to 40 %) obtained in the country. It is worth noting that the spatial distribution of soil organic matter corresponds to low to high soil total nitrogen, available phosphorus, soil pH, exchangeable bases, etc. According to Žížala et al. (2021), the SOC of the Czech Republic ranged from 0.08 to 3.99 % at 0–30 cm. However, the dominant SOC content in the country ranged from 1–1.5 %. The value is higher than what is obtained in Nigeria but similar to Morocco because of comparable climatic characteristics. Similarly, the pH value of the country ranged from 2.9 to 7.5 at a depth of 0–30 cm.

Soil organic carbon (SOC) accumulation is sensitive to land use, farming practices, and environmental factors (Xie et al., 2021). However, Hengl et al. (2015) reported elevation as the most critical influence in the spatial prediction of SOC in topsoil Africa. Similarly, Wang et al. (2012) observed that elevation, slope, soil clay, and water contents explained the variability of SOC and N in Western Australia. Conversely, Žížala et al. (2021) reported that terrain attributes explain SOC variability at a coarser resolution than a more detailed resolution. Terrain and bioclimatic variables were reported to control the spatial distribution of soil water status, dynamics of plant litter mineralization, as well as erosion and deposition processes (Hengl et al., 2015). Furthermore, the function of elevation in explaining spatial variability of SOC and N, for instance, can be linked to corresponding differences in soil temperature and the intensity of cultivation, which is stronger in lower locations due to accessibility than in higher positions.

## **1.6 Sampling strategies and sampling ratio**

The purpose of soil sampling is to evaluate a field's soil nutrient status and properties as accurately as possible while considering the associated financial budget (Dinkins and Jones, 2008). Also, soil sampling strategies assess the soil nutrient status from which fertilization, liming and other soil management recommendations can be given (Flowers et al., 2005). In addition, the spatial variability of soil nutrients between and within regions can result in surplus and limited fertilization. Soil sampling could be traditional or more detailed (Dawson and Knowles, 2018). Traditional soil sampling considers sites as homogenous areas of similar soil nutrients and

properties distribution (Flowers et al., 2005). In contrast, detailed sampling strategies such as grid or directed soil sampling aim to look at soil nutrients and properties spatially to improve soil production through a proper soil management approach. Sampling is the most expensive part of a survey since it involves principles like survey intensity, geographic variability, and mapping scale (Webster and Oliver, 1992).

Numerous farming operations still use traditional sampling strategies as the best way to sample fields. In this sampling strategy, a composite soil sample is obtained by randomly probing various locations across the sampling region and then combining them into one sample. Although the method may be inexpensive, it has the substantial drawback of inadequately characterizing field variability, resulting in coarse maps with clear, defined borders separating sampled regions (Crozier and Heiniger, 2001).

Detailed sampling strategies (e.g., regular systematic sampling) (called cell centre sampling) involve taking one sample from the centre of each grid cell (Flowers et al., 2005). Studies have shown that the closer the sample point distance, the more accurate the assumptions and predictive models between the soil nutrients and the sites under investigation (Wollenhaupt, 1994). However, Franzen and Peck (1995) recommended that grid sampling points be decided by the uniformity of the field, soil types, past management and financial budget. Besides that soil sampling strategies, some studies have focused on estimating the variogram of soil properties (Minasny and McBratney, 2006; Vašát et al., 2010). Nevertheless, sampling on a point grid (square, triangular, or hexagonal) is generally advised for consistently shaped regions when no auxiliary information is provided. In these cases, the equilateral triangular grid offers the most accurate assessment of the desired attribute (Yfantis et al., 1987). According to Webster and Oliver (1992), at least 150 observations are required to estimate the variogram correctly.

On the other hand, many regions of interest are irregularly shaped, necessitating a more complex sampling strategies optimization approach. One optimization method provides an ideal distribution of sample locations in a geographical area via environmental attributes. The space-filling or spatial coverage sample is the term for this approach (Royle and Nychka, 1998). It makes the distance between each sample place as short as possible. However, determining the appropriate sampling strategies for mapping with machine learning models has not yet been considered in detail in digital soil mapping research.

According to Boettinger et al. (2010), simple random, stratified random, and regular sampling strategies are most suitable for statistical inference since they introduce randomness at the early

stage of the sampling strategy. Apart from the sampling strategies mentioned above, other soil sampling strategies, model-based sampling strategies, and conditioned Latin hypercube sampling (cLHS) exist (Boettinger et al., 2010). In model-based sampling procedures, the spatial properties of the soil samples and the auxiliary factors that impact pedogenesis are considered (Sun et al., 2017; Thomas et al., 2015). Meanwhile, cLHS incorporates continuous and categorical environmental covariates into the sampling strategy (Minasny and McBratney, 2006). Thanks to advances in remote sensing technology, soil covariate factors, such as vegetation maps, geology maps, and their derivatives, may now be employed as auxiliary variables for digital soil mapping (Higo et al., 2015; Brus and De Gruijter, 1997; Brungard and Boettinger, 2010). In addition, these datasets are essential in influencing the spatial distribution of soil parameters, and representative ancillary variables can be selected for model-based soil sampling to determine the best geographical areas for soil samples (Viscarra Rossel et al., 2010). Minasny and McBratney (2006) employed a cLHS sampling approach to building one optimal and efficient soil sample plan using a digital elevation model, slope, compound topographic index, and normalized difference vegetation index as supplementary data. Qin et al. (2011) used cluster analysis of soil environmental variables to find representative sample spots and then developed a sampling design approach. These researchers quickly created a realistic and appropriate sampling strategy by obtaining extensive ancillary information. As a result, environmental variables, particularly in a field or local regions, cannot provide correct supplemental information for the spatial unbiasedness of soil attributes (Ciampalini et al., 2015).

Soil sample ratio or size is the number of observations that reflects the spatial variability of a target soil property in a given space (McBratney and Webster, 1983; Adetunji, 1994). By maximizing the sample ratio and recognizing representative sampling areas, field sampling, a significant problem in soil surveys, helps acquire reliable soil mapping data (Wang et al., 2012). Sampling strategy and sample ratio are the most vital criteria to consider in predicting soil parameters that vary spatially based on the vast heterogeneity of soil environments (McBratney and Webster, 1983). In DSM, appropriate sampling strategies and sample ratios remain challenging (Lai et al., 2021). Both sampling strategy and sample ratio limit most modelling regime accuracy. Some sampling schemes may require many observation points to make accurate soil nutrient recommendations in highly complex sites (for example, grid neural sampling scheme). Unfortunately, even though this method may present the reality of the soil condition, it is labour-

intensive (for instance, the time taken to collect soil at a large scale) and involves enormous financial implications for laboratory analysis (Higo et al., 2015).

Soil sampling and sample ratio are the initial step in obtaining site-specific information to make liming, fertilization, and other soil management decisions. Choosing the right sampling strategy ensures that the soil in a field is gathered in a way that yields the most accurate and dependable soil test findings. Wang et al. (2019) proposed a sampling strategy with good spatial coverage and feature space coverage for precise farm field-level soil mapping for a limited sample ratio. Use a sampling approach that best captures that variance since soils in agricultural fields might vary greatly (Pal et al., 2009). When using a site-specific management method, proper sampling is very critical. However, where financial limitations exist, soil samples are collected to represent important agricultural soils, essential geology information, toposquence order, pedological soil information, and mineral weathering sequence. In addition, a sequence of soil types is recognized to isolate the effects of a single factor as far as possible. These approaches have been adopted at local, regional, and national scales (Pal et al., 2009). Besides that, the ideal sample size varies greatly depending on the soil type, sampling depth, and nutrient type (Adetunji, 1994). According to Adetunji (1994), 25 to 30 and 30 to 40 core samples are advised in the tropical region for newly opened ground and highly farmed land, respectively. Conversely, Minasny and McBratney (2006) opined that optimal sample size accurately represents the variability in the environmental covariates and provides enough samples for predictive models.

## 2. AREAS OF STUDY

The research presented in this thesis was conducted in three different countries with diverse environmental and soil conditions. The studies were conducted in Nigeria, Morocco and Czech Republic (Figure 1).

Nigeria, with a land area of 923,768 km<sup>2</sup>, is in West Africa, between longitudes 3° and 14° and latitudes 3 ° and 14° (Awala et al., 2019). Nigeria has three different climatic zones: a tropical monsoon climate in the south, a tropical savannah climate in the centre and northern areas, and a Sahelian hot and semi-arid climate in the north (Amanchukwu et al., 2015). Soil conditions are characterized by low inherent soil fertility, as evidenced by low soil organic matter, base saturation and low CEC (Esu, 2005). The country is actively engaged in agriculture with high agriculture production and intensity.

Morocco is surrounded by the North Atlantic Ocean and the Mediterranean Sea in the northwest corner of Africa, with a total land area of 446,550 km<sup>2</sup> (Schilling et al., 2012). The Atlantic Ocean to the west, the Mediterranean Sea to the north, and the Sahara Desert to the south and southeast significantly impact rainfall and temperature. Between October and May, the country receives most of its rain with an average annual rainfall of 1,200 mm, with temperatures ranging from 18°C to 28°C in the summer and 8°C to 17°C in the winter (Tuel et al., 2021). Soil conditions of the country are characterized by high pH value, high potassium content and relatively high soil organic matter (Lahmar et al., 2020).

The Czech Republic is a landlocked country in the middle region of Europe. The country has a total of 78,864 km<sup>2</sup> of land area, with agricultural lands representing more than 50% of the total area of the Czech Republic (Sklenicka, 2006). The country is in a transition zone, with a climate influenced by maritime and continental air masses (Hradecký and Brázdil, 2016). The soil profile of the Czech Republic consists of some rich, black chernozems and good-quality brown soils in the drier and lower areas.



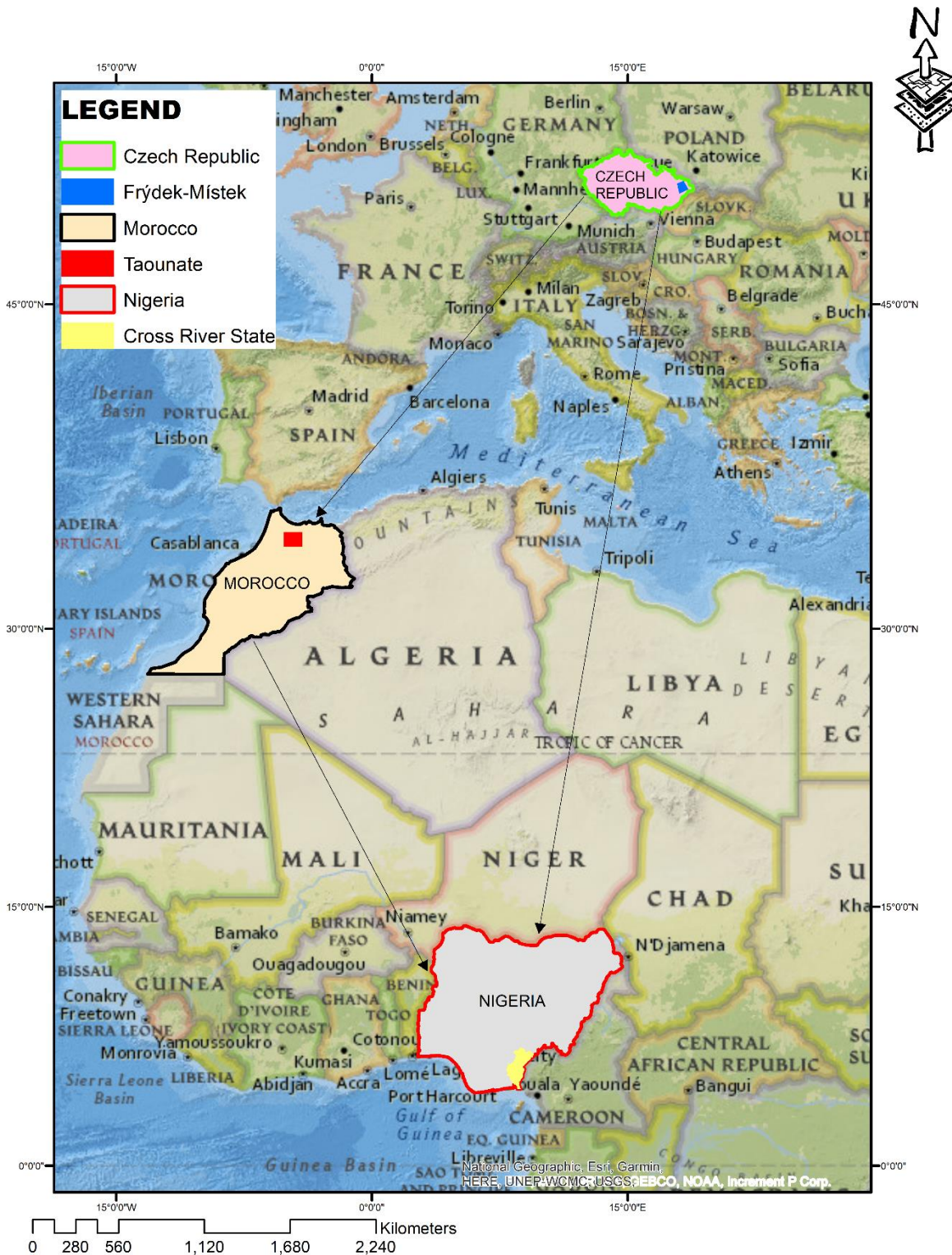


Figure 1. Map of the countries where the studies were conducted showing the specific regions of sampling.

## 2.1 Cross River State, Nigeria

Figure 1 is the map of the Cross River State in the southeastern region of Nigeria, showing the study area of some parts of the thesis. In the research work conducted in Calabar, Akamkpa, Yakurr and Ogoja Local Government Areas, soil sampling was conducted using a simple sampling strategy. This is due to the field having similar variability due to natural soil qualities (e.g., soil texture and drainage) and soil management history (e.g., drainage and previous land use).

Calabar extends from latitudes 4°51'45.86"N and longitude 8°19'50.69"E and spreads over an area of approximately 200 km<sup>2</sup> with an elevation of about 44 m above sea level (John et al., 2018). It is under the humid tropical rainforest zones, marked by two distinct seasons (rainy and dry). The area receives the mean annual rainfall above 3500 mm with a temperature range between 22°C and 30°C and relative humidity of 83 % (John et al., 2018). The mainland uses include rain-fed cultivation of tree crops and arable crops. The dominant landscape units in the study area are relatively flat terrain. The soils of the study area are developed from a coastal plain sand parent material (Akpan-Idiok et al., 2012; Afu et al., 2019). They are characterized by udic moisture and isohyperthermic temperature regimes, respectively (Soil Survey Staff, 2014). In addition, according to USDA soil taxonomic classification, the soil order of the region is overwhelmingly Ultisols, and the soil is classified as Typic Kandiodults (Soil Survey Staff, 2014). In the area, unconsolidated materials occurred with high sand and silt content (Akpan-Idiok et al., 2012; Afu et al., 2019). Due to the lithological homogeneity observed by the high sand content, the soil class distribution in the landscape is similar, with a high occurrence of Inceptisols and Ultisols (Esu, 2005).

The Ogoja area is covered by the southern guinea savannah and cultivated for oil palm, teak, maize, sugar cane, cassava, groundnut, oil palm, vegetable crops and paddy rice. At the same time, tropical rainforests surround the Yakurr and Akamkpa areas. Yakurr and Akamkpa have similar climates and vegetation and often experience slight temperature variations. Temperature varies from 23 to 34 °C in the Ogoja area and 23 to 32 °C in Yakurr and Akamkpa areas (Sambo et al., 2016). A sub-humid tropical climate with distinct wet and dry seasons characterizes the region. Rainfall is between 1500 and 2500 mm per year, and relative humidity is between 80 and 90 %. They are characterized by udic moisture and isohyperthermic temperature regimes, respectively (Soil Survey Staff 2014).

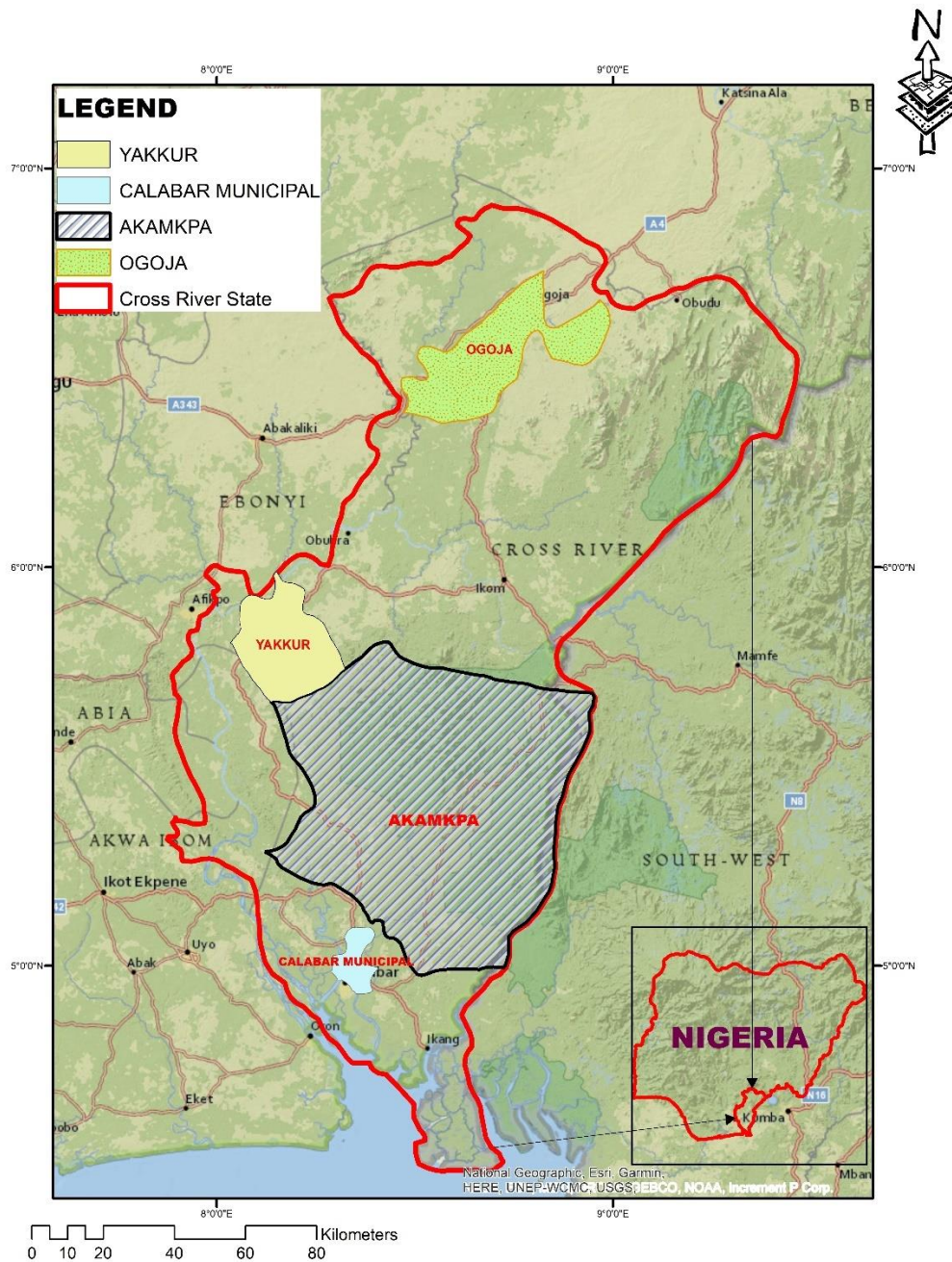


Figure 2. Shows the map of Cross River State in southeastern Nigeria with different study areas. Calabar Municipal Area is closer to the coastal zone; Akamkpa Local Government Area is predominantly the forested area of the state; Yakurr and Ogoja Local Government Areas are composed of grassland and forest areas.

## 2.2 Frýdek-Místek district, Czech Republic

The research was conducted in the Moravian-Silesian Region's foothills, Frýdek-Místek district in the Czech Republic (Figure 3), to study soil sulphur variability. The site is an active agricultural site situated at geographical coordinates of latitude  $49^{\circ} 41' 0''$  N and longitude  $18^{\circ} 20' 0''$  E and the elevation of 225–327 m above sea level. Meanwhile, according to the Koppen classification system, the area's climate is classified as Cfb = Temperate oceanic climate with high rainfall even in dry months. The study area is approximately  $889.8 \text{ km}^2$  designated for agricultural activities with scattered trees. The area's soils are characterized by a cambic diagnostic horizon that distinguishes them with a fine sandy loam texture. The soil contains more than 4% clay concentration and a lithic discontinuity with reduced carbonate content (Kozak et al., 2010).

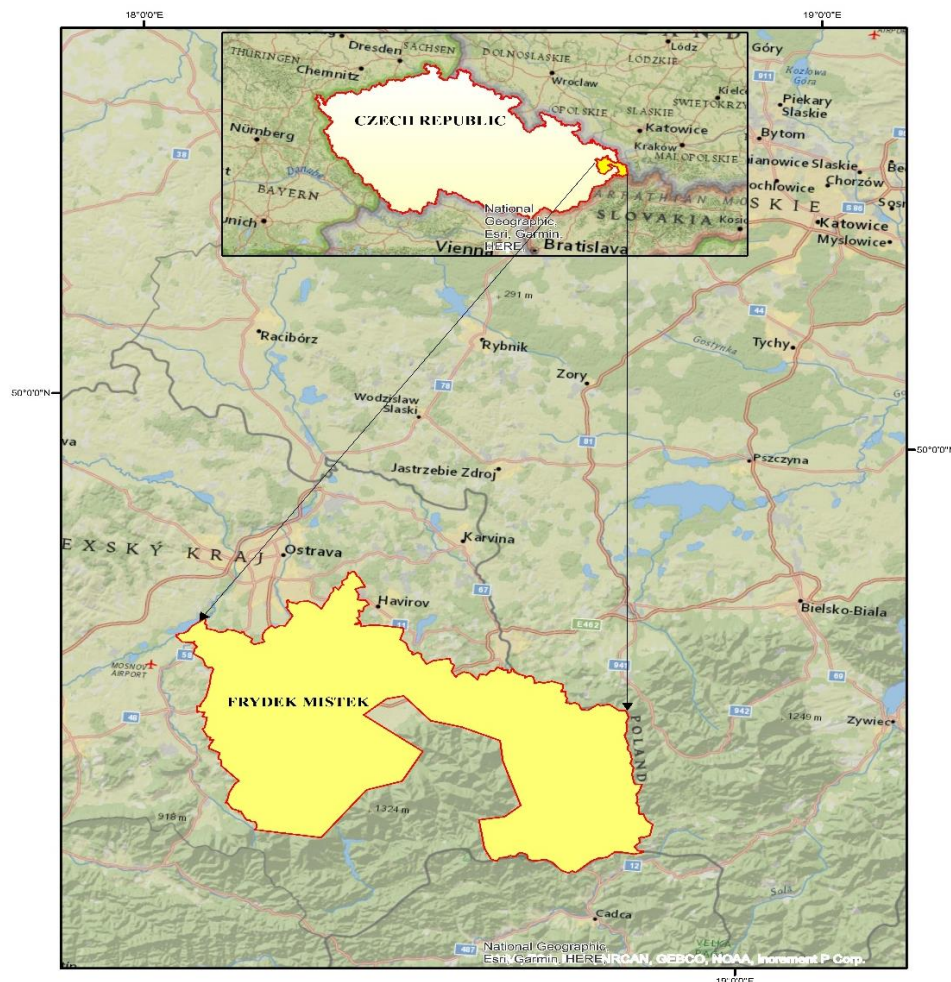


Figure 3. Map of Frydek Mistek situated in the Moravian-Silesian Region's foothills of the Czech Republic

### **2.3 Taounate province, Morocco**

We also conducted some studies in Morocco, testing sample ratios and sampling schemes in predicting soil phosphorus and potassium. The systematic method was applied by subdividing the study area using a regular 1 km square grid to select sampling points while ensuring complete coverage of the study area (Figure 4). A total of (n = 1470) samples were collected to a depth of 0–40 cm over two months (November and December 2013). Figure 3, the selected area for the study belonging to Taounate province in the north of Morocco (34° 47'N, 4° 4.4'W and 34° 05'N, 5° 10.3'W), is displayed as a rectangle of 7979 km<sup>2</sup> (101 km x 79 km) (Fig. 1). Jbel Oudka is the most important mountain of the Taounate region, with an elevation that reaches 1587 m and is characterized by significant vegetation cover. The area covered by this study contains a part of the Atlas Mountains in the northwest. In general, the altitude ranges from 78 to 1969 m. A Mediterranean climate and irregular rainfall characterize this region. According to the Köppen-Geiger classification, the study area is in the CSA class with a mean annual temperature of 17.8 °C and mean precipitation of 549 mm. As a result, the average maximum temperature of the hottest month is approximately 34.2°C and the average minimum of the coldest month is 0.5°C (Allali et al., 2020; Rezouki et al., 2021). Geological formations of the Taounate wrinkle consist of a Jurassic-Cretaceous series of marl overcome molassic formations composed of sandstone and conglomerates (Mesrar et al., 2017).

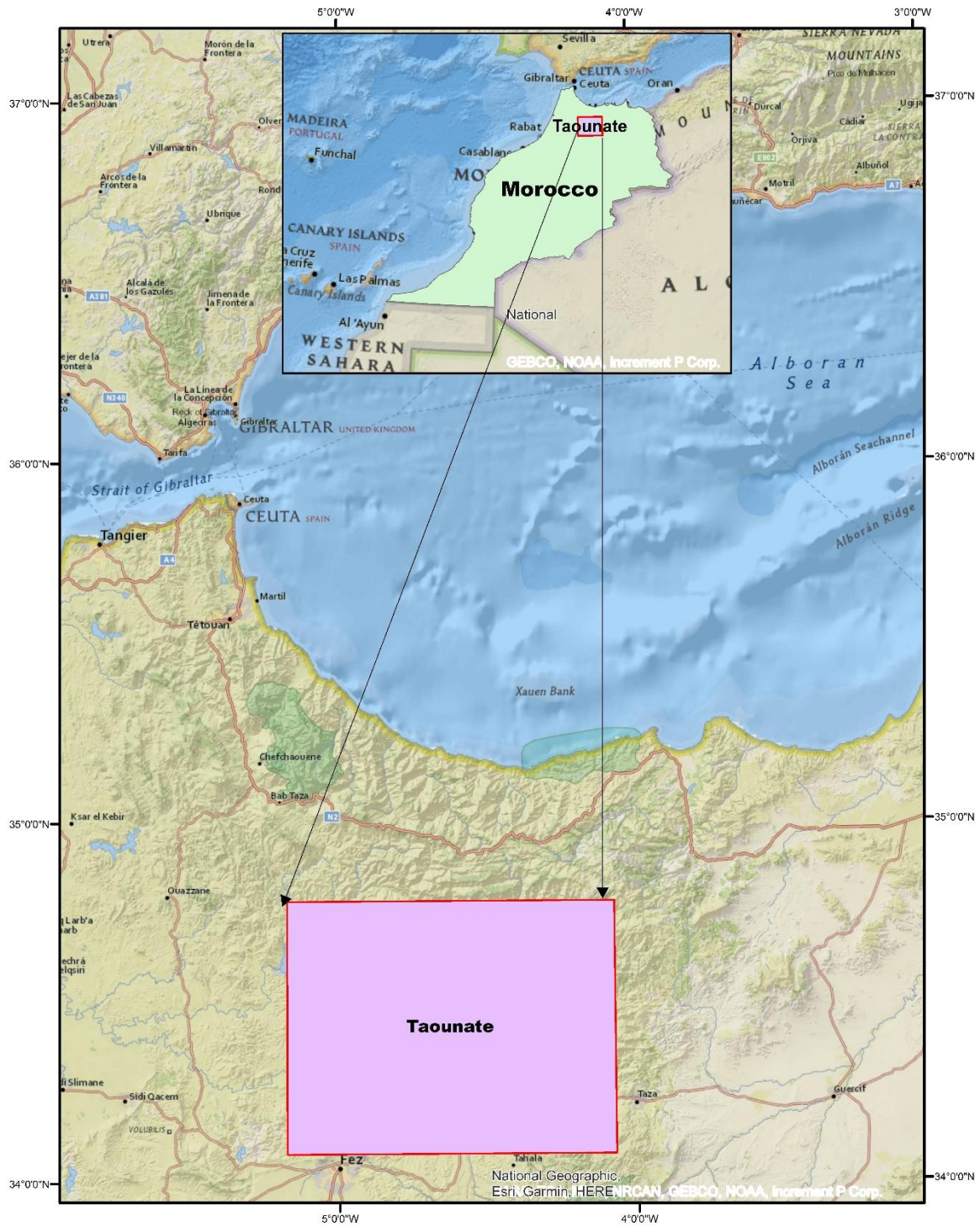


Figure 4. Map of study area showing training and testing sample points at a depth of 0–40 cm. The study area is by the coast of the Mediterranean.

### 3. AIMS AND HYPOTHESES

Testing the DSM approach's performance for site-specific management necessitated this study. Crop and underlying soil management occur at a smaller scale than the whole field. Therefore, it is essential to test the workability of the DSM technique for site-specific management using few samples, variation of sample ratios and sampling strategies, and environmental covariates. These investigations may lead to the administration of soil fertilizer and lime rates depending on soil laboratory analysis results, and plant demands particular to that location, maximizing total field production—site-specific management results in a variable rate of applied amendments. The quantity of amendments needed in a field varies depending on the soil nutrient status (soil testing), the crop produced, soil texture, drainage, and landscape position.

The study's general aim is to test different DSM methods, sampling strategies and ratios to estimate soil nutrients and properties. Digital soil mapping technique is still developing in Nigeria and Morocco. Therefore, it is essential to test and show the technique's usefulness in developing updated soil spatial variability of soil nutrients and properties.

#### 3.1 Aims

1. Testing different geostatistics and machine learning models to estimate soil nutrients and properties with different environmental covariates.

This objective was achieved in the following research articles:

**Article 1:** John, K., S. M. Afu, I. A. Isong, E. E. Aki, N. M. Kebonye, E. O. Ayito, P. A. Chapman, M. O. Eyong, and V. Penížek. "Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics." *International Journal of Environmental Science and Technology* (2021): 1-16.

**Article 2:** John, K., Afu, S. M., Isong, I. A., Chapman, P. A., Kebonye, N. M., & Ayito, E. O. "Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the Southeastern soils of Nigeria." *Environmental Engineering & Management Journal (EEMJ)*, 20(7).

**Article 3:** John, K., Agyeman, P. C., Kebonye, N. M., Isong, I. A., Ayito, E. O., Ofem, K. I., & Qin, C. Z. "Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur." *Catena*, 206, 105534.

2. Testing different multiple linear modelling approaches in estimating soil organic matter with some selected soil physical properties.

This objective was achieved in the research article below:

**Paper 4:** Ofem, K.I., John, K., Pawlett, M., Eyong, M.O., Awaogu, C.E., Umeugokwe, P., Ambrose-Igho, G., Ezeaku, P.I. and Asadu, C.L.A., 2021. Estimating Soil Organic Matter: A Case Study of Soil Physical Properties for Environment-Related Issues in Southeast Nigeria. *Earth Systems and Environment*, 5(4), 899-908.

3. Evaluating the role of sampling and sample ratios in a digital soil mapping approach of soil nutrients.

This objective was achieved in the article below:

**Article 5:** John, K., Bouslihim, Y., Bouasria, A., Razouk, R., Hssaini, L., Isong, I. A., Ayito, E. O., Ambrose-Igho, G. (2022). Assessing the impact of sampling strategy in Random Forest-based predicting of soil nutrients: a study case from Northern Morocco *Geocarto International* (2022): 1-14.

### **3.2 Hypotheses**

The hypotheses of the study are as follows:

1. Applying more robust geostatistics and machine learning models may be helpful in mapping soil properties and soil nutrients.
2. Sampling strategies and sample ratios may contribute to accurately mapping soil nutrients.



## **4. SYNTHESIS AND CONCLUDING REMARKS**

### **4.1 Synthesis of key findings**

In summary, the present thesis work is a compilation of published articles. The thesis synthesized different findings on digital soil mapping techniques of soil properties and nutrients performed under other sampling strategies with varying sample ratios.

In the studies conducted in Nigeria, under a similar sampling approach (e.g., simple random sampling strategies), we tested different predictive models (e.g., empirical Bayesian kriging-EBK, MLR, ordinary kriging-OK, cokriging-Cok and inverse distance weighting-IDW). These simple models were employed to demonstrate their performance and how they could be improved with an increase in soil sample ratios and sampling strategies.

Firstly, in the Calabar Municipal region of Nigeria, we tested a novel geostatistic model [e.g., empirical Bayesian kriging (EBK)] with a few samples obtained on a plain surface and the results compared with the simple multiple regression model. In exploring the dataset, we observed via principal component analysis (PCA) that soil properties and environmental covariates (terrain derivatives) explained 78.1 % of the variability in the dataset. After then, we applied the EBK model, which is more robust than OK and uses the autocorrelation function by simulating the few samples obtained. The interpolation output presented good predictions for Mg, K, P, pH, and TN ( $R^2 \geq 0.5$ ). At the same time, the linear regression model performed poorly except in the case of Mg, where a good prediction was obtained only. In addition, in this particular area where this study was conducted, although characterized by relative plain relief, we observed that soil properties variability might not be explained more accurately with the terrain derivatives at a field scale. However, the study proved that EBK is more robust, utilizing a few sample points to explain the spatial distribution of soil properties and nutrients.

Further in the study, in the Akamkpa region of Nigeria, SOC was mapped only using ordinary kriging (OK,) cokriging (Cok) and inverse distance weighting (IDW) models. A slightly undulating landscape characterizes the area where this study was carried out. With the different models applied in the study, SOC demonstrated a moderate spatial dependence and explained the importance of estimating SOC spatial variability. Using the Cok model, terrain derivatives were incorporated to improve the spatial structure of the SOC variability. We observed that the Cok produced a smaller mean error due to adding terrain derivatives. Cok prediction via the significantly correlated terrain attributes (elevation, LS factor, and profile curvature) improved the

map structure compared to OK and IDW. The Cok map was more detailed, showing the capability of terrain attributes to be robust ancillary variables for improving detailed spatial SOC maps.

Since the Cok model showed prospects in mapping SOC, we further applied it to develop a more robust hybrid model for mapping soil sulphur, a vital soil nutrient, with samples collected via grid sampling schemes. The study tested cokriging (Cok) and the ML technique (Gaussian process regression; GPR) and then compared their performance with a hybrid machine learning model called a cokriging-Gaussian process regression (Cok-GPR) model under grid sampling strategies. The hybrid method used the Cok matrices to predict soil sulphur via GPR. All parameters, sulphur (S), calcium (Ca), potassium (K), magnesium (Mg), sodium (Na), phosphorus (P), and vanadium (V), were estimated via inductively coupled plasma optical emission spectroscopy (ICP-OES) equipment. 80% of the datasets were used for calibration, while the remaining 20% were used for validation. In Cok, we used all the data for cross-validation, and the results showed that Cok1 (Ca, K, Na) performed better than Cok2 (P and Mg) and Cok3 (V). In the GPR models, GPR1 (Ca, K, Na) performed better than GPR2 and GPR3. However, in the Cok-GPR models, all models were generally improved and were within acceptable model criteria—using a Taylor diagram, Cok1-GPR outperformed Cok and GPR models, respectively.

The fourth article in the study set out to test different linear model functions to minimize variables' collinearity in soil organic matter (SOM) prediction in southeastern Nigeria in randomly collected samples. This study pioneered a novel technique to understand SOM variation in soils across different sedimentary lithologies (e.g., limestone, Shale–limestone-sandstone intercalation, alluvium and sandstone-limestone intercalation). This study tested various multiple linear regression functions on soils developed on different sedimentary lithologies to explain the inter-relationship between SOM and bulk density (BD), saturated hydraulic conductivity (Ksat), total (Total P), air-filled (Air P), and capillary porosities (Cap P). The study revealed a strong relationship between SOM and Ksat, BD in soils developed on limestone. After then, we applied the different linear regression functions in the leap package in R for SOM estimation, and the best linear regression function was leapbackward (RMSE = 11.50 %,  $R^2 = 0.58$ , MAE = 8.48 %), which produced a smaller error when compared with leapforward, leapseq, and lmStepAIC functions.

The study fifthly set out to explain the spatial sensitivity of sample ratios with sampling strategies in soil nutrient prediction. The findings obtained in the study revealed that random sampling was suitable for predicting phosphorus, whereas the conditioned Latin hypercube sampling (cLHS)

was suitable for predicting potassium. Furthermore, model accuracy improved when the sample ratio increased in both random sampling and cLHS for phosphorus and potassium prediction. This is because an increase in sample ratio may improve the predicting accuracy and somewhat offset the influence of inappropriate sampling strategies. This finding may be attributed to the fact that sampling strategies largely depend on the understanding, knowledge, and experience of the spatial variation of soil nutrients and properties. And similar findings were obtained in sampling strategies studies by Zhao et al. (2016).

## **4.2 Concluding remarks**

The findings in the thesis revealed the pragmatic nature of EBK as a model able to maximize a small sample ratio via a simple sampling strategy to produce high accuracy compared to multivariate statistics. The multivariate statistics (e.g., PCA) revealed that selected study soil properties and nutrients exhibit a strong relationship compared to the terrain derivatives in a low relief condition.

Comparing the performance of the OK model with Cok and IDW in the mapping of SOC (article 2), we observed that even with a 50 % increase in samples (compared to article 1), Cok and IDW outperformed OK as the samples were not sufficient to establish a strong autocorrelation function with OK via a simple sampling strategy (article 2). Generally, the study showed that the EBK model has higher accuracy and lower uncertainty with fewer samples (article 1). However, it would be interesting to learn whether increasing the samples may improve the accuracy of EBK, as in the case of Cok for SOC mapping (article 2).

On the other hand, applying a novel Cok-GPR model showed some prospects in accurately estimating soil sulphur. Hence, it is inferred that combining geostatistics and ML could be an exciting aspect of exploring soil nutrient mapping (article 3).

In this thesis, we pointed out that increasing the sample ratio with corresponding random or conditioned LHS sampling strategies in a random forest model can also provide more insight into accurately estimating soil nutrients in a given landscape (article 5).

Specific conclusions of the different studies are presented below:

Under a simple random sampling strategy, comparing the performance of geostatistic and deterministic models in SOC prediction, Cok and IDW performed well. Therefore, soil and land users could adopt SOC maps by Cok and IDW, as the maps revealed low SOC in the study area. These maps will be a vital tool in planning the different nutrient needs of crops for adequate agricultural production productivity. In addition, the created maps could be used as a reference

point for various soil purposes, ranging from sampling optimization to updating soil maps with more ancillary variables. The research sets a precedent for future digital soil mapping in Nigeria. Therefore, future studies should include other auxiliary data with a more robust model and cover a broader range of soil types to improve model performance.

Under the grid sampling scheme, we mapped soil sulphur via a proposed cokriging-Gaussian process regression. The model more precisely showed the spatial distribution of soil sulphur (S) levels in the actively cultivated agricultural soil. The Cok-GPR model had higher fitting accuracy and robustness than Cok and GPR models. Even though Cok-GPR has a higher computational cost, it yielded the best prediction. Therefore, the proposed Cok-GPR model may be applied to efficiently predict soil nutrient element levels, the products used for proper soil fertilization calculations, and precise soil management practices.

The thesis found that using some selected soil physical parameters and parent material to predict soil organic matter (SOM) via different linear functions could be helpful in understanding soil interaction. The best-performing model function in the study was leapbackward, which yielded a lesser error when compared to the other linear functions. Furthermore, all models identified bulk density and hydraulic conductivity as the most critical variables in explaining SOM variation across different sedimentary geologies.

Under two different sampling schemes (e.g., random and cLHS) and two sample ratios, we predicted potassium and phosphorus via a random forest model. The adopted approach showed some prospects for precisely and accurately predicting soil nutrients in the Mediterranean region. The conditioned Latin hypercubes and random sampling techniques in predicting soil nutrient levels exhibited success and robustness. For phosphorus prediction, random sampling worked well, while conditioned Latin hypercubes sampling worked well for potassium. According to the findings, both sampling schemes are susceptible to increased sample ratios.

In conclusion, the thesis presented relevant and specific strategies to develop accurate estimates of different soil properties ranging from site situation, sample ratio, predictive models, and sampling strategies. We concluded that while digital soil mapping is rapidly approaching the point where it can meet various soil information demands, challenges such as new theories, techniques, and applications of digital soil mapping still exist. And this must be addressed in the future, particularly for highly plain and heterogeneous relief and human-affected environments, optimum sample ratios and sampling strategies. Therefore, we recommend that attention be given to the sample

strategies and sampling ratio with corresponding more robust models for accurately estimating soil properties and nutrients.

## REFERENCES

- Adetunji, M. T. (1994). Optimum sample size and sampling depth for soil nutrient analysis of some tropical soils. *Communications in soil science and plant analysis*, 25(3-4), 199-205.
- Agyeman, P. C., Kebonye, N. M., John, K., Borůvka, L., Vašát, R., & Fajemisim, O. (2022). Prediction of nickel concentration in peri-urban and urban soils using hybridized empirical bayesian kriging and support vector machine regression. *Scientific Reports*, 12(1), 1-16.
- Afu, S. M., Isong, I. A. & Awaogu, C. E (2019). Agricultural potentials of floodplain soils with contrasting parent material in Cross River State, Nigeria. *Global Journal of Pure and Applied Science*, 25, 13-22.
- Akpan-Idiok, A. U., Ogbaji, P. O., & Antigha, N. R. B. (2012). Infiltration, degradation rate and vulnerability potential of Onwu river floodplain soils in Cross River State, Nigeria. *Journal of Agriculture, Biotechnology and Ecology*, 5(1), 62-74.
- Al Masmoudi, Y., Bouslihim, Y., Doumali, K., El Aissaoui, A., & Namr, K. I. (2021). Application of the random forest model to predict the plasticity state of vertisols. *Journal of Ecological Engineering*, 22(2).
- Allali, A., Rezouki, S., Lougraimzi, H., Touati, N., Eloutassi, N., & Fadli, M. (2020). Agricultural traditional practices and risks of using insecticides during seed storage in morocco. *Plant Cell Biotechnol. Mol. Biol*, 29-37.
- Amanchukwu, R. N., Amadi-Ali, T. G., & Ololube, N. P. (2015). Climate change education in Nigeria: The role of curriculum review. *Education*, 5(3), 71-79.
- Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., & Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spatial Statistics*, 14, 91-113.
- Arrouays, D., Lagacherie, P., & Hartemink, A. E. (2017). Digital soil mapping across the globe. *Geoderma Regional*, 9, 1-4.
- Awala, F. O., Ndukwu, B. C., & Agbagwa, I. O. (2019). Phytogeographical Distribution and Fruit Diversity of *Lagenaria siceraria* Species in Nigeria. *American Journal of Plant Sciences*, 10(6), 958-975.
- Bardgett, R. D., & Wardle, D. A. (2010). Aboveground-belowground linkages: biotic interactions, ecosystem processes, and global change. Oxford University Press.
- Bangroo, S. A., Najar, G. R., Achin, E., & Truong, P. N. (2020). Application of predictor variables in spatial quantification of soil organic carbon and total nitrogen using regression kriging in the North Kashmir forest Himalayas. *Catena*, 193, 104632.

- Beaudette, D. E., & O'Geen, A. T. (2009). Soil-Web: an online soil survey for California, Arizona, and Nevada. *Computers & Geosciences*, 35(10), 2119-2128.
- Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E. D., & Goldschmitt, M. (2005). Digital soil mapping using artificial neural networks. *Journal of plant nutrition and soil science*, 168(1), 21-33.
- Behrens, T., Zhu, A. X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155(3-4), 175-185.
- Bell, J. C., Cunningham, R. L., & Havens, M. W. (1992). Calibration and validation of a soil-landscape model for predicting soil drainage class. *Soil Science Society of America Journal*, 56(6), 1860-1866.
- Bell, J. C., Cunningham, R. L., & Havens, M. W. (1994). Soil drainage class probability mapping using a soil-landscape model. *Soil Science Society of America Journal*, 58(2), 464-470.
- Bolstad, P. V., & Stowe, T. (1994). An evaluation of DEM accuracy: elevation, slope, and aspect. *Photogrammetric Engineering & Remote Sensing*, 60(11), 1327-1332.
- Bregt, A. K. (1992). Processing of soil survey data. Wageningen University and Research. <https://www.proquest.com/openview/c1b4e1a18021c03a7bcd4c923356a357/1?pq-origsite=gscholar&cbl=18750&diss=y>.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Brungard, C. W., & Boettinger, J. L. (2010). Conditioned latin hypercube sampling: Optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In *Digital soil mapping* (pp. 67-75). Springer, Dordrecht.
- Brungard, C., Nauman, T., Duniway, M., Veblen, K., Nehring, K., White, D., Salley, S., & Anchang, J. (2021). Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma*, 397, 114998.
- Brus, D. J., & De Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1-2), 1-44.
- Bui, E. N., & Moran, C. J. (2003). A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. *Geoderma*, 111(1-2), 21-44.
- Bui, E. N. (2021). Machine learning in the Australian critical zone. In *Data Science Applied to Sustainability Analysis* (pp. 43-78). Elsevier.

Bui, E. N., Henderson, B. L., & Viergever, K. (2006). Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling*, 191(3-4), 431-446.

Burrough, P., & McDonnell, R. (1998). Optimal interpolation using geostatistics in *Principles of Geographical Information Systems* Oxford.

Cappello, C., De Iaco, S., Palma, M., & Pellegrino, D. (2021). Spatio-temporal modeling of an environmental trivariate vector combining air and soil measurements from Ireland. *Spatial Statistics*, 42, 100455.

Carré, F., McBratney, A. B., & Minasny, B. (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, 141(1-2), 1-14.

Chang, K. T., & Tsai, B. W. (1991). The effect of DEM resolution on slope and aspect mapping. *Cartography and geographic information systems*, 18(1), 69-77.

Chen, S., Mulder, V. L., Martin, M. P., Walter, C., Lacoste, M., Richer-de-Forges, A. C., Saby, N.P., Loiseau, T., Hu, B. & Arrouays, D. (2019). Probability mapping of soil thickness by random survival forest at a national scale. *Geoderma*, 344, 184-194.

Chen, Z., Chen, D., Zhao, C., Kwan, M.P., Cai, J., Zhuang, Y., Zhao, B., Wang, X., Chen, B., Yang, J. & Li, R., 2020. Influence of meteorological conditions on PM<sub>2.5</sub> concentrations across China: A review of methodology and mechanism. *Environment international*, 139, 105558.

Ciampalini, A., André, F., Garfagnoli, F., Grandjean, G., Lambot, S., Chiarantini, L., & Moretti, S. (2015). Improved estimation of soil clay content by the fusion of remote hyperspectral and proximal geophysical sensing. *Journal of Applied Geophysics*, 116, 135-145.

Connolly, J., Holden, N. M., & Ward, S. M. (2007). Mapping peatlands in Ireland using a rule-based methodology and digital data. *Soil Science Society of America Journal*, 71(2), 492-499.

Cressie, N., & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 209-226.

Crozier, C. R., & R. W. Heiniger. (2001). Soil Facts: Soil Sampling for Precision Farming Systems." North Carolina Cooperative Extension Service, North Carolina State University. <http://www.soil.ncsu.edu/publications/Soilfacts/AG-439-36/AG-439-36.pdf> (last accessed 12 August 2022).

Davatgar, N., Neishabouri, M. R., & Sepaskhah, A. R. (2012). Delineation of site specific nutrient management zones for a paddy cultivated area based on soil fertility using fuzzy clustering. *Geoderma*, 173, 111-118.

Dawson, A., & Knowles, O. (2018). To grid or not to grid—a review of soil sampling strategies. Farm environmental planning—science, policy and practice. Fertilizer and Lime Research Centre, Massey University, Palmerston North. Available at: <http://flrc.massey.ac.nz/publications.html>.



- Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), 1-13.
- Dietrich, W. E., Reiss, R., Hsu, M. L., & Montgomery, D. R. (1995). A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. *Hydrological processes*, 9(3-4), 383-400.
- Dinkins, C. P., & C. Jones. (2008). *Soil Sampling Strategies*. Department of Land Resources and Environmental Sciences, Montana State University. [http://landresources.montana.edu/soilfertility/PDFbyformat/publication%20pdfs/Soil\\_Sampling\\_Strat\\_MT200803AG.pdf](http://landresources.montana.edu/soilfertility/PDFbyformat/publication%20pdfs/Soil_Sampling_Strat_MT200803AG.pdf) (last accessed 22 August 2022).
- Esu, I. E. (2005). Characterization, classification and management problems of the major orders in Nigeria. 26th Inaugural Lecture, Department of Soil Science University of Calabar. 38-59.
- Fanuel, L., Kibebew, K., Tekalign, M., & Hailu, S. (2018). Accounting spatial variability of soil properties and mapping fertilizer types using geostatistics in Southern Ethiopia. *Communications in Soil Science and Plant Analysis*, 49(1), 124-137.
- Ferguson, R. B., Hergert, G. W., Schepers, J. S., Gotway, C. A., Cahoon, J. E., & Peterson, T. A. (2002). Site-specific nitrogen management of irrigated maize: Yield and soil residual nitrate effects. *Soil Science Society of America Journal*, 66(2), 544-553.
- Florinsky, I. V. (1998). Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, 12(1), 47-62.
- Flowers, M., Weisz, R., & White, J. G. (2005). Yield-based management zones and grid sampling strategies: Describing soil test and nutrient variability. *Agronomy Journal*, 97(3), 968-982.
- Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PloS one*, 12(1), e0170478.
- Franzen, D. W., & Peck, T. R. (1995). Field soil sampling density for variable rate fertilization. *Journal of Production Agriculture*, 8(4), 568-574.
- Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., & Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: The Cook Agronomy Farm data set. *Spatial Statistics*, 14, 70-90.
- Ge, F., Zhang, J., Su, Z., & Nie, X. (2007). Response of changes in soil nutrients to soil erosion on a purple soil of cultivated sloping land. *Acta Ecologica Sinica*, 27(2), 459-463.
- Gessler, P. E., Moore, I. D., McKenzie, N. J., & Ryan, P. J. (1995). Soil-landscape modelling and spatial prediction of soil attributes. *International journal of geographical information systems*, 9(4), 421-432.

- Giasson, E., Ten Caten, A., Bagatini, T., & Bonfatti, B. (2015). Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. *Ciência Rural*, 45, 1592-1598.
- Giustini, F., Ciotoli, G., Rinaldini, A., Ruggiero, L., & Voltaggio, M. (2019). Mapping the geogenic radon potential and radon risk by using Empirical Bayesian Kriging regression: A case study from a volcanic area of central Italy. *Science of the Total Environment*, 661, 449-464.
- Goovaerts, P. (2011). A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties. *European journal of soil science*, 62(3), 371-380.
- Gotway, C. A., Ferguson, R. B., Hergert, G. W., & Peterson, T. A. (1996). Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, 60(4), 1237-1247.
- Gribov, A., & Krivoruchko, K. (2020). Empirical Bayesian kriging implementation and usage. *Science of the Total Environment*, 722, 137290.
- 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.
- Grunwald, S., Thompson, J. A., & Boettinger, J. L. (2011). Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal*, 75(4), 1201-1213.
- Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, 152(3-4), 195-207.
- Guo, P. T., Li, M. F., Luo, W., Tang, Q. F., Liu, Z. W., & Lin, Z. M. (2015). Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma*, 237, 49-59.
- Hartemink, A. E., McBratney, A., & Mendonça-Santos, M. (2008). Digital soil mapping: a state of the art. *Digital Soil Mapping with Limited Data*. Springer, Netherlands, 3-14.
- Heisel, T., Ersbøll, A. K., & Andreasen, C. (1999). Weed mapping with co-kriging using soil properties. *Precision Agriculture*, 1(1), 39-52.
- Haileselassie, B., Stomph, T. J., & Hoffland, E. (2011). Teff (*Eragrostis tef*) production constraints on Vertisols in Ethiopia: farmers' perceptions and evaluation of low soil zinc as yield-limiting factor. *Soil Science and Plant Nutrition*, 57(4), 587-596.
- Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. (2005). Australia-wide predictions of soil properties using decision trees. *Geoderma*, 124(3-4), 383-398.
- Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2), 75-93.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., 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., Miller, M.A., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M. & McGrath, S.P. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, 11(1), 1-18.

Heisel, T., Ersbøll, A. K., & Andreasen, C. (1999). Weed mapping with co-kriging using soil properties. *Precision Agriculture*, 1(1), 39-52.

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

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, 265, 62-77.

Higo, M., Isobe, K., Yamaguchi, M., & Torigoe, Y. (2015). Impact of a soil sampling strategy on the spatial distribution and diversity of arbuscular mycorrhizal communities at a small scale in two winter cover crop rotational systems. *Annals of Microbiology*, 65(2), 985-993.

Hooshmand, A., Delgh, M., Izadi, A., & Aali, K. A. (2011). Application of kriging and cokriging in spatial estimation of groundwater quality parameters. *African Journal of Agricultural Research*, 6(14), 3402-3408.

Hradecký, J., & Brázdil, R. (2016). Climate in the past and present in the Czech lands in the Central European Context. In *Landscapes and Landforms of the Czech Republic* (pp. 19-28). Springer, Cham.

Hudson, B. D. (1992). The soil survey as paradigm-based science. *Soil Science Society of America Journal*, 56(3), 836-84.

Hussain, I., Shakeel, M., Faisal, M., Soomro, Z. A., Hussain, M., & Hussain, T. (2014). Distribution of total dissolved solids in drinking water by means of bayesian kriging and gaussian spatial predictive process. *Water Quality, Exposure and Health*, 6(4), 177-185.

Iqbal, J., Thomasson, J. A., Jenkins, J. N., Owens, P. R., & Whisler, F. D. (2005). Spatial variability analysis of soil physical properties of alluvial soils. *Soil Science Society of America Journal*, 69(4), 1338–1350.

James G, Witten D, Hastie T, Tibshirani R (2014) *An introduction to statistical learning: with applications in R*. Springer Publishing Company Incorporated, Berlin

Jenny, H. (1941). *Factors of soil formation: a system of quantitative pedology*. New York: McGraw-Hill.

John, K., Ayito, E. O., & Odey, S. (2018). Interaction between some soil physicochemical properties and weather variables on sub-humid tropical rainforest soils of Cross River State, Southeastern Nigeria. *Annual Research & Review in Biology*, 1-12.

John, K., Lawani, S. O., Esther, A. O., Ndiye, K. M., Sunday, O. J., & Penížek, V. (2019). Predictive Mapping of Soil Properties for Precision Agriculture Using Geographic Information System (GIS) Based Geostatistics Models. *Modern Applied Science*, 13(10).

John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., & Marcus Afu, S. (2020). Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. *Land*, 9(12), 487.

John, K., Kebonye, N. M., Agyeman, P. C., & Ahado, S. K. (2021). Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data. *Environmental Monitoring and Assessment*, 193(4), 1-15.

John, K., Afu, S. M., Isong, I. A., Aki, E. E., Kebonye, N. M., Ayito, E. O., Chapman, P. A., Eyong, M. O., Penížek, V. (2021). Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. *International Journal of Environmental Science and Technology*, 18(11), 3327-3342.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. New York, NY: Academic Press Inc.

Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O. & Borůvka, L. (2021). Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma*, 384, p.114792.

Kempen, B., Heuvelink, G. B. M., Brus, D. J., & Stoorvogel, J. J. (2010). Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. *European Journal of Soil Science*, 61(3), 333-347.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B. M., & de Vries, F. (2012). Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. *Soil Science Society of America Journal*, 76(6), 2097.

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81, 401-418.

Kidd, D., Searle, R., Grundy, M., McBratney, A., Robinson, N., O'Brien, L., Zund, P., Arrouays, D., Thomas, 420 M., & Padarian, J. (2020). Operationalising digital soil mapping—Lessons from Australia. *Geoderma* 421 Regional, e00335.

- Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3), 226-239.
- Kozák, J., Němeček, J., Borůvka, L., Kodešová, R., Janků, J., Jacko, K. and Hladík, J. (2010). *Soil Atlas of the Czech Republic*, Czech University of Life Sciences, Prague, 150 pp
- Krivoruchko, K., & Gribov, A. (2019). Evaluation of empirical Bayesian kriging. *Spatial Statistics*, 32, 100368.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28, 1-26.
- Lagacherie, P., & McBratney, A. B. (2006). Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. *Developments in soil science*, 31, 3-22.
- Lagacherie, P., & McBratney, A. B. (2007). Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: *Digital Soil Mapping: An Introductory Perspective* (eds. P. Lagacherie, A. B. McBratney & M. Voltz), pp. 3–24. Elsevier, Amsterdam, The Netherlands.
- Lahmar, M., El Khodrani, N., Omrania, S., Dakak, H., Moussadek, R., Douaik, A., Iaaich, H., & Zouahri, A. (2020). Assessment of the Quality of Soil and Groundwater of the Agricultural Area of Sidi Yahya Region, Morocco. In *E3S Web of Conferences* (Vol. 150, p. 01001). EDP Sciences.
- Lai, Y. Q., Wang, H. L., & Sun, X. L. (2021). A comparison of importance of modelling method and sample size for mapping soil organic matter in Guangdong, China. *Ecological Indicators*, 126, 107618.
- Lal, R. (2015). Restoring soil quality to mitigate soil degradation. *Sustainability*, 7(5), 5875-5895.
- Lark, R. M., & Beckett, P. H. T. (1998). A geostatistical descriptor of the spatial distribution of soil classes, and its use in predicting the purity of possible soil map units. *Geoderma*, 83(3-4), 243-267.
- Lark, R. M., & Webster, R. (2006). Geostatistical mapping of geomorphic variables in the presence of trend. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, 31(7), 862-874.
- Li, G., Messina, J. P., Peter, B. G., & Snapp, S. S. (2017). Mapping land suitability for agriculture in Malawi. *Land degradation & development*, 28(7), 2001-2016.
- Li, Y., Hernandez, J. H., Aviles, M., Knappett, P. S. K., Giardino, J. R., Miranda, R., Puy, M. J., Padilla, F., & Morales, J. (2020). Empirical Bayesian Kriging method to evaluate inter-annual water-table evolution in the Cuenca Alta del Río Laja aquifer, Guanajuato, México. *Journal of Hydrology*, 582, 124517.

- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.
- Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., & Viscarra Rossel, R. (2019). Baseline map of soil organic matter in China and its associated uncertainty. *Geoderma*, 335, 47-56.
- Lilburne, L. R., Hewitt, A. E., & Webb, T. W. (2012). Soil and informatics science combine to develop S-map: A new generation soil information system for New Zealand. *Geoderma*, 170, 232-238.
- Lima, C. H. R., Kwon, H.-H., & Kim, Y.-T. (2021). A Bayesian Kriging model applied for spatial downscaling of daily rainfall from GCMs. *Journal of Hydrology*, 597, 126095.
- Ma, Y., Minasny, B., & Wu, C. (2017). Mapping key soil properties to support agricultural production in Eastern China. *Geoderma Regional*, 10, 144-153.
- MacMillan, R. A., Moon, D. E., & Coupé, R. A. (2007). Automated predictive ecological mapping in a forest region of BC, Canada, 2001–2005. *Geoderma*, 140(4), 353-373.
- Malone, B. P., Jha, S. K., Minasny, B., & McBratney, A. B. (2016). Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma*, 262, 243-253.
- Matheron, G. (1969). *Le krigeage universel* (Vol. 1). Paris: École nationale supérieure des mines de Paris.
- McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1-2), 3-52.
- McKenzie, N. J. & Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma* 89: 67–94.
- McKenzie, N. J., Gessler, P. E., Ryan, P. J., & O'Connell, D. A. (2000). The role of terrain analysis in soil mapping. In 'Terrain analysis: principles and applications'.(Eds JP Wilson, JC Gallant) pp. 245–265.
- McSweeney, K., Slater, B. K., David Hammer, R., Bell, J. C., Gessler, P. E., & Petersen, G. W. (1994). Towards a new framework for modeling the soil-landscape continuum. Factors of soil formation: A fiftieth anniversary retrospective, 33, 127-145.
- Mesrar, H., Sadiki, A., Faleh, A., Quijano, L., Gaspar, L., & Navas, A., 2017. Vertical and lateral distribution of fallout <sup>137</sup>Cs and soil properties along representative toposequences of central Rif, Morocco. *J. Environ. Radioact.*, 169, 27-39.

- Messina, J. P., Peter, B. G., & Snapp, S. S. (2017). Re-evaluating the Malawian farm input subsidy programme. *Nature plants*, 3(4), 1-9.
- Metcalf, H., Milne, A. E., Webster, R., Lark, R. M., Murdoch, A. J., & Storkey, J. (2016). Designing a sampling scheme to reveal correlations between weeds and soil properties at multiple spatial scales. *Weed Research*, 56(1), 1-13.
- Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & geosciences*, 32(9), 1378-1388.
- McBratney, A. B., & Webster, R. (1983). How many observations are needed for regional estimation of soil properties?. *Soil Science*, 135(3), 177-183.
- Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, 264, 301-311.
- Mirzaei, R., & Sakizadeh, M. (2015). Comparison of interpolation methods for the estimation of groundwater contamination in Andimeshk-Shush Plain, Southwest of Iran. *Environmental Science and Pollution Research*, 23(3), 2758-2769.
- Mishra, U., Lal, R., Liu, D., & Van Meirvenne, M. (2010). Predicting the spatial variation of the soil organic carbon pool at a regional scale. *Soil Science Society of America Journal*, 74(3), 906-914.
- Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil science society of america journal*, 57(2), 443-452.
- Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E., & Mehnatkesh, A. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environmental monitoring and assessment*, 188(3), 195.
- Mpeketula, P. M. G. (2016). Soil organic carbon dynamics and mycorrhizal fungal diversity in contrasting agroecosystems. Michigan State University. [https://d.lib.msu.edu/etd/3907/datastream/OBJ/download/Soil\\_Organic\\_Carbon\\_Dynamics\\_and\\_Mycorrhizal\\_Fungal\\_Diversity\\_in\\_Contrasting\\_Agroecosystems.pdf](https://d.lib.msu.edu/etd/3907/datastream/OBJ/download/Soil_Organic_Carbon_Dynamics_and_Mycorrhizal_Fungal_Diversity_in_Contrasting_Agroecosystems.pdf).
- Mueller, T. G., & Pierce, F. J. (2003). Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. *Soil Science Society of America Journal*, 67(1), 258-267.
- Mulder, V. L., De Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping—A review. *Geoderma*, 162(1-2), 1-19.
- Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., & Arrouays, D. (2016). GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. *Science of the Total Environment*, 573, 1352-1369.

- Nabiollahi, K., Taghizadeh-Mehrjardi, R., Shahabi, A., Heung, B., Amirian-Chakan, A., Davari, M., & Scholten, T. (2021). Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. *Geoderma*, 385, 114858.
- Nafiu, A. K., Abiodun, M. O., Okpara, I. M., & Chude, V. O. (2012). Soil fertility evaluation: a potential tool for predicting fertilizer requirement for crops in Nigeria. *African Journal of Agricultural Research*, 7(47), 6204-6214.
- Nussbaum, M., Papritz, A., Baltensweiler, A., & Walthert, L. (2014). Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. *Geoscientific Model Development*, 7(3), 1197-1210.
- Nyéki, A., Kerepesi, C., Daróczy, B., Benczúr, A., Milics, G., Nagy, J., Harsányi, E., Kovács, A.J., & Neményi, M. (2021). Application of Spatio-temporal data in site-specific maize yield prediction with machine learning methods. *Precision Agriculture*, 22(5), 1397-1415.
- Odeh, I. O., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3-4), 215-226.
- Odgers, N. P., McBratney, A. B., & Minasny, B. (2011). Bottom-up digital soil mapping. I. Soil layer classes. *Geoderma*, 163(1-2), 38-44.
- Pal, D. K., Bhattacharyya, T., Srivastava, P., Chandran, P., & Ray, S. K. (2009). Soils of the Indo-Gangetic Plains: their historical perspective and management. *Current Science*, 1193-1202.
- Panday, D., Maharjan, B., Chalise, D., Shrestha, R. K., & Twanabasu, B. (2018). Digital soil mapping in the Bara district of Nepal using kriging tool in ArcGIS. *PloS one*, 13(10), e0206350.
- Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Scientific reports*, 9(1), 1-12.
- Pásztor, L., Laborczi, A., Takács, K., Illés, G., Szabó, J., & Szatmári, G. (2020). Progress in the elaboration of GSM conform DSM products and their functional utilization in Hungary. *Geoderma Regional*, 21, e00269.
- Penížek, V., & Borůvka, L. (2006). Soil depth prediction supported by primary terrain attributes: a comparison of methods. *Plant, Soil and Environment*, 52(9), 424-430.
- Penížek, V., Zádorová, T., Kodešová, R., & Vaněk, A. (2016). Influence of elevation data resolution on spatial prediction of colluvial soils in a Luvisol region. *Plos one*, 11(11), e0165699.
- Pilz, J., & Spöck, G. (2008). Why do we need and how should we implement Bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, 22(5), 621-632.



- Qin, C. Z., Zhu, A. X., Pei, T., Li, B. L., Scholten, T., Behrens, T., & Zhou, C. H. (2011). An approach to computing topographic wetness index based on maximum downslope gradient. *Precision agriculture*, 12(1), 32-43.
- Quinlan, J. R. (1992). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343-348).
- R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. [online]. Available at <https://www.r-project.org/>. (Verified 13 May 2020).
- Rezouki, S., Allali, A., Louasté, B., Eloutassi, N., & Fadli, M., 2021. Physico-chemical evaluation of soil resources in different regions of Taza–Taounate, Morocco. *Mediterr. J. Chem.*, 11(1), 1-9.
- Robinson, T. P., & Metternicht, G. (2003). A comparison of inverse distance weighting and ordinary kriging for characterising within-paddock spatial variability of soil properties in Western Australia. *Cartography*, 32(1), 11-24.
- Rossel, R. V., Chen, C., Grundy, M. J., Searle, R., Clifford, D., & Campbell, P. H. (2015). The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. *Soil Research*, 53(8), 845-864.
- Royle, J. A., & Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, 24(5), 479-488.
- Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.
- Sambo, E. E., Ufoegbune, G. C., Eruola, A. O. & Ojekunle, O. Z. (2016). Impact of Rainfall Variability on Flooding of Rivers in Cross River Basin, Nigeria. Nigerian Metreological Society (NMETS), "Climate Variability And Change: Impact, Science, Innovation And Policy" at Federal College of Education Osiele, Abeokuta. 21—24 November, 2016
- Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34-38.
- Schabenberger, O., & Gotway, C. A. (2017). *Statistical methods for spatial data analysis: Texts in statistical science*. Chapman and Hall/CRC.
- Schilling, J., Freier, K. P., Hertig, E., & Scheffran, J. (2012). Climate change, vulnerability and adaptation in North Africa with focus on Morocco. *Agriculture, Ecosystems & Environment*, 156, 12-26.
- Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D. (2003). Predictive soil mapping: a review. *Progress in Physical geography*, 27(2), 171-197.

- Searle, R., McBratney, A., Grundy, M., Kidd, D., Malone, B., Arrouays, D., Stockman, U., Zund, P., Wilson, P., Wilford, J. & Van Gool, D. (2021). Digital soil mapping and assessment for Australia and beyond: A propitious future. *Geoderma Regional*, 24, e00359.
- Shaddad, S. M., Madrau, S., Castrignanò, A., & Mouazen, A. M. (2016). Data fusion techniques for delineation of site-specific management zones in a field in UK. *Precision agriculture*, 17(2), 200-217.
- Singh, A., Santra, P., Kumar, M., Panwar, N., & Meghwal, P. R. (2016). Spatial assessment of soil organic carbon and physicochemical properties in a horticultural orchard at arid zone of India using geostatistical approaches. *Environmental monitoring and assessment*, 188(9), 1-19.
- Sklenicka, P. (2006). Applying evaluation criteria for the land consolidation effect to three contrasting study areas in the Czech Republic. *Land use policy*, 23(4), 502-510.
- Soil Survey Staff (2014). *Keys to Soil Taxonomy*, 12th Edn Washington. DC: Natural Resources Conservation Service, United States Department of Agriculture.
- Steinbuch, L., Brus, D. J., & Heuvelink, G. B. (2022). Mapping depth to Pleistocene sand with Bayesian generalized linear geostatistical models. *European Journal of Soil Science*, 73(1), e13140.
- Sun, X. L., Wang, H. L., Zhao, Y. G., Zhang, C., & Zhang, G. L. (2017). Digital soil mapping based on wavelet decomposed components of environmental covariates. *Geoderma*, 303, 118-132.
- 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, 115153.
- Takata, Y., Funakawa, S., Akshalov, K., Ishida, N., & Kosaki, T. (2007). Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. *Soil science and plant nutrition*, 53(3), 289-299.
- Theobald, D. (1989). Accuracy and bias issues in surface representation. In *The accuracy of spatial databases* (pp. 77-82). CRC Press.
- Thomas, M., Clifford, D., Bartley, R., Philip, S., Brough, D., Gregory, L., Willis, R. & Glover, M. (2015). Putting regional digital soil mapping into practice in Tropical Northern Australia. *Geoderma*, 241, 145-157.
- Tuel, A., Kang, S., & Eltahir, E. A. (2021). Understanding climate change over the southwestern Mediterranean using high-resolution simulations. *Climate Dynamics*, 56(3), 985-1001.
- Tziachris, P., Metaxa, E., Papadopoulos, F., & Papadopoulou, M. (2017). Spatial modelling and prediction assessment of soil iron using kriging interpolation with pH as auxiliary information. *ISPRS International Journal of Geo-Information*, 6(9), 283.

- Vašát, R., Heuvelink, G. B. M., & Borůvka, L. (2010). Sampling design optimization for multivariate soil mapping. *Geoderma*, 155(3-4), 147-153.
- Viscarra Rossel, R. A., Rizzo, R., Demattê, J. A. M., & Behrens, T. (2010). Spatial Modeling of a Soil Fertility Index using Visible–Near-Infrared Spectra and Terrain Attributes. *Soil science society of America journal*, 74(4), 1293-1300.
- Wang, S., Wang, X., & Ouyang, Z. (2012). Effects of land use, climate, topography and soil properties on regional soil organic carbon and total nitrogen in the Upstream Watershed of Miyun Reservoir, North China. *Journal of Environmental Sciences*, 24(3), 387-395.
- Wang, Y., Jiang, L., Qi, Q., Liu, Y., & Wang, J. (2019). Remote Sensing-Guided Sampling Design with Both Good Spatial Coverage and Feature Space Coverage for Accurate Farm Field-Level Soil Mapping. *Remote Sensing*, 11(16), 1946.
- Walton, J. T. (2008). Subpixel urban land cover estimation. *Photogrammetric Engineering & Remote Sensing*, 74(10), 1213-1222.
- Webster, R., & Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of soil science*, 43(1), 177-192.
- Webster, R., & Oliver, M. A. (2001). *Geostatistics for environmental scientists*. John Wiley & Sons.
- Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afrotropical landscape. *Ecological Indicators*, 52, 394-403.
- Wilson, J. P., & Gallant, J. C. (2000). Digital terrain analysis. *Terrain analysis: Principles and applications*, 6(12), 1-27.
- Wong, D. W. (2016). Interpolation: Inverse-Distance Weighting. *International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology*, 1-7.
- Wollenhaupt, N. C., Wolkowski, R. P., & Clayton, M. K. (1994). Mapping soil test phosphorus and potassium for variable-rate fertilizer application. *Journal of production agriculture*, 7(4), 441-448.
- Wu, Z., Yi, L., & Zhang, G. (2009). Uncertainty analysis of object location in multi-source remote sensing imagery classification. *International journal of remote sensing*, 30(20), 5473-5487.
- Xie, E., Zhang, Y., Huang, B., Zhao, Y., Shi, X., Hu, W., & Qu, M. (2021). Spatiotemporal variations in soil organic carbon and their drivers in southeastern China during 1981-2011. *Soil and Tillage Research*, 205, 104763.
- Xue, Y., Liu, Y., Ji, C., Xue, G., & Huang, S. (2020). System identification of ship dynamic model based on Gaussian process regression with input noise. *Ocean Engineering*, 216, 107862.

Yang, L., Zhu, A. X., Qi, F., Qin, C. Z., Li, B., & Pei, T. (2013). An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. *International Journal of Geographical Information Science*, 27(1), 1-23.

Yfantis, E. A., Flatman, G. T., & Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology*, 19(3), 183-205.

Zádorová, T., Žížala, D., Penížek, V., & Vaněk, A. (2020). Harmonisation of a large-scale historical database with the actual Czech soil classification system. *Soil and Water Research*, 15(2), 101-115.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M.B., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R. and Xu, M. (2020). Conventional and digital soil mapping in Iran: Past, present, and future. *Catena*, 188, 104424.

Zhang, L., Yang, L., Ma, T., Shen, F., Cai, Y., & Zhou, C. (2021). A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. *Geoderma*, 384, 114809.

Zhao, Y., Xu, X., Tian, K., Huang, B., & Hai, N. (2016). Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. *Environmental Earth Sciences*, 75(1), 1-14.

Zhu, A.X., (1997). A similarity model for representing soil spatial information. *Geoderma* 77, 217 – 242.

Zhu, A.X. (2000). Mapping soil-landscape as spatial continua: the neural network approach. *Water Resources Research* 36, 663 – 677.

Žížala, D., Juřicová, A., Kapička, J., & Novotný, I. (2021). The potential risk of combined effects of water and tillage erosion on the agricultural landscape in Czechia. *Journal of Maps*, 17(2), 428-438.

## PUBLICATION LIST

1. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics

**John, K.**, S. M. Afu, I. A. Isong, E. E. Aki, N. M. Kebonye, E. O. Ayito, P. A. Chapman, M. O. Eyong, and V. Penížek. "Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics." *International Journal of Environmental Science and Technology* (2021): 1-16.

2. Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the southeastern soils of Nigeria

**John, K.**, Afu, S. M., Isong, I. A., Chapman, P. A., Kebonye, N. M., & Ayito, E. O. "Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the Southeastern soils of Nigeria." *Environmental Engineering & Management Journal (EEMJ)*, 20(7).

3. Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur.

**John, K.**, Agyeman, P. C., Kebonye, N. M., Isong, I. A., Ayito, E. O., Ofem, K. I., & Qin, C. Z. "Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur." *Catena*, 206, 105534.

4. Estimating Soil Organic Matter: A Case Study of Soil Physical Properties for Environment-Related Issues in Southeast Nigeria

Ofem, K.I., **John, K.**, Pawlett, M., Eyong, M.O., Awaogu, C.E., Umeugokwe, P., Ambrose-Igho, G., Ezeaku, P.I. and Asadu, C.L.A., 2021. Estimating Soil Organic Matter: A Case Study of Soil Physical Properties for Environment-Related Issues in Southeast Nigeria. *Earth Systems and Environment*, 5(4), 899-908.

4. Assessing the impact of sampling strategy in Random Forest-based predicting of soil nutrients: a study case from Northern Morocco

**John, K.**, Bouslihim, Y., Bouasria, A., Razouk, R., Hssaini, L., Isong, I. A., Ayito, E. O., Ambrose-Igho, G. (2022). Assessing the impact of sampling strategy in Random Forest-based predicting of soil nutrients: a study case from Northern Morocco. *Geocarto International* (2022): 1-14.



# Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics

K. John<sup>1</sup> · S. M. Afu<sup>2</sup> · I. A. Isong<sup>2</sup> · E. E. Aki<sup>2</sup> · N. M. Kebonye<sup>1</sup> · E. O. Ayito<sup>2</sup> · P. A. Chapman<sup>1</sup> · M. O. Eyong<sup>2</sup> · V. Penížek<sup>1</sup>

Received: 27 June 2020 / Revised: 19 September 2020 / Accepted: 16 December 2020 / Published online: 4 January 2021  
© Islamic Azad University (IAU) 2021

## Abstract

The spatial modelling of soil properties provides us with essential and useful information relevant to soil fertility management and environmental protection. The study aims to investigate the ability of empirical Bayesian kriging and principal component analysis, multiple linear regressions with environmental covariates in the modelling of soil properties distribution. For this study, thirty ( $n = 30$ ) soil samples were obtained at 0–30 cm depth and nine (9) soil-environmental covariates derived from the digital elevation model (Shutter Radar Topography Mission at 30 m resolution) in southeastern Nigeria. The summary statistics revealed high sand content ( $> 70\%$ ) which revealed that the soils of the humid tropics developed on the coastal plain parent material are coarse-textured. Pearson correlation matrix revealed a significant but weak correlation between soil properties and soil-environmental variables. Using empirical Bayesian kriging interpolation, the cross-validation results revealed an acceptable prediction for magnesium, potassium, phosphorus, pH and total nitrogen ( $R^2 > 0.5$  with RMSE closer to 0). The principal component analysis reveals that principal component 1 to principal component 5 could interpret 78.1% of the total variability of soil properties. Modelling each soil property using multiple linear regression with the derived soil-environmental covariates, the study noted that only magnesium gave the best model fit with 50.9% of the soil-environmental covariates explaining its variability, while other soil properties presented unacceptable models. Therefore, to improve soil property prediction through multiple linear regression, more observation points are recommended to interpret better the performance of multiple linear regression over flat terrain system.

**Keywords** Empirical Bayesian kriging · Humid tropical soils · Multiple linear regression · Principal component analysis · Terrain attributes

## Introduction

The soils of humid tropical Africa are highly weathered soils and are similar to those in the Amazon, dominated by Alfisols, Ultisols, Oxisol and Inceptisols soil orders (Soil Survey Staff 2014). These soils are exposed to high weathering

conditions. For example, the soils in South-East Nigeria receive excessive amounts of precipitation which amount to over 3500 mm per annum (Esu 2005). In the humid tropics, precipitation exceeds evaporation in the condition that the soils are developed. The soils are generally characterized by low inherent fertility which is an indication of low soil organic carbon, total nitrogen content, activity clays and exchangeable cations (Sanchez 1977; Akpan-Idiok 2012; Delarmelinda et al. 2017; John et al. 2018, 2019). These soil properties are conditioned by rainfall patterns, vegetation distribution, parent material, topography, vegetation, time (Jenny 1941) and other soil-environmental covariates which Bishop and McBratney (2001) and Zhang et al. (2017) refer to it as a subset of the soil-forming factors.

The application of the soil-environmental covariates over the years in soil mapping has been successful. Climate and terrain derivatives are amongst the most widely used

Editorial responsibility: Samareh Mirkia.

✉ K. John  
johnk@af.czu.cz

<sup>1</sup> Department of Soil Science and Soil Protection, Faculty of Agrobiolgy, Food, and Natural Resources, Czech University of Life Sciences, Kamýcká 129, 165 00 Praha-Suchdol, Prague, Czech Republic

<sup>2</sup> Department of Soil Science, Faculty of Agriculture, University of Calabar, P.M.B. 1115, Calabar, Nigeria



environment covariates in soil property modelling (Zhang et al. 2017). Climatic data are obtained from the interpolation of temperature and precipitation from meteorological stations to more advanced remote sensing data acquisition, surface soil moisture, temperature and evapotranspiration (Boettinger et al. 2008; Petropoulos et al. 2015), while terrain derivatives are obtained either from topogrid maps or open-source digital elevation model satellite platforms. Other covariates such as normalized difference vegetation index (NDVI), soil maps and landcover maps have also been successfully employed in soil property modelling (Zhu et al. 2010; Qin et al. 2012).

In soil property mapping, different multivariate statistics such as generalized linear model (GLM), principal component analysis (PCA), structural equation models (SEM), generalized additive models (GAM), random forest (RF), support vector machine (SVM), artificial neural network (ANN) in conjunction with different interpolation techniques such as ordinary kriging (OK), regression kriging (RK), inverse distance weighting (IDW) and most recently, the empirical Bayesian kriging (EBK) have been used and are yielding good results (Bishop and McBratney 2001; Esu et al. 2014; Odeh et al. 2006; Beguería et al. 2013; de Carvalho Junior et al. 2014; Zeraatpisheh et al. 2019; Gribov and Krivoruchko 2020). These methods of modelling have been adopted to evaluate the accurate spatial distribution of soil property with soil-environment covariates. However, the combinations of different statistical and interpolation methods have led to the progress made in advanced soil mapping. Zeraatpisheh et al. (2019) reported the combination of inverse distance weighting interpolation and clustered analysis in the study of soils in the semi-arid region of Iran. According to their studies, soil organic carbon distribution was spatially explained by vegetation index covariate using the clustered analysis and principal components. These advanced methods of mapping try to explain the relationship between soil properties and environmental covariates (Beguería et al. 2013; Park and Vlek 2002; de Carvalho Junior et al. 2014; Akpa et al. 2014).

In as much as progress has been made in soil property mapping, little or no effort has been made in the application of multivariate statistics and geostatistics in the mapping of soil properties formed on a relatively flat terrain condition in the humid tropical rainforest. Therefore, the study hypothesizes that in relatively flat terrain system, soil-environmental covariates may contribute little or nothing in soil properties variability.

This present study tries to model soil properties distribution at the local scale level using a probability kriging method called EBK. The EBK was applied because it is a very robust and reliable interpolation for both automatic and interactive data interpolation. EBK comprises of two geostatistical models: the intrinsic random function kriging (Chilès

and Delfiner 1999; Gribov and Krivoruchko 2020) and linear mixed model that is called kriging with an external trend in digital soil mapping researches (Varentsov et al. 2020). EBK exemplifies a logical and viable statistical technique that could be applied to explain soil property distribution as a method to expose soil nutrient-deficiency, limiting crop productivity in the region. Further details on EBK are presented by Gribov and Krivoruchko (2020). According to Hussain et al. (2014), EBK is most suitable for spatial prediction of total dissolved solids (TSD) in drinking water. Also, Mirzaei and Sakizadeh (2016) reported that EBK model was more superior to other interpolation techniques such as OK and IDW for estimation of groundwater contamination. More so, since the area experiences seasonal flooding due to the high amount of rainfall and poor drainage condition, it was necessary to adopt the EBK model for the study. Therefore, this study aims to investigate the ability of EBK and PCA, multiple linear regression (MLR) with environmental covariates in the modelling of soil properties distribution in flat terrain system.

This research is carried out at the Department of Soil Science, University of Calabar, Nigeria and Department of Soil Science and Soil Protection of Czech University of Life Sciences, Prague, Czech Republic from January to September 2019.

## Materials and methods

### Basic idea

#### First step

The EBK technique, which is a geostatistical technique available in ArcGIS Desktop (Release 10.7 Redlands, CA: Environmental Systems Research Institute) was used to map soil property distribution in the study. The EBK method is a more practical geostatistical technique compared to other forms of kriging methods (Krivoruchko and Gribov 2014). This technique accommodates related uncertainties in plotting the semivariogram and automates the most difficult aspects of composing an adequate kriging model (Samsonova et al. 2017). The principles governing the technique includes the interpolation of a mapped property to any specific point (pixel), the variogram model is estimated from the data, and at each of the input data locations, a new value is simulated which then generates a new semivariogram model estimated from the simulated data using the Bayesian rule (Eq. 1),

$$P(X, Y) = P(X|Y)P(Y) = P(Y, X) = P(Y|X)P(X) \quad (1)$$



$P(X, Y)$  is known as the posterior, parameter to be estimated.  $P(Y|X)$  is referred to as the likelihood of an event.  $P(X)$  is referred to as the prior.  $P(Y)$  is called the marginal likelihood, and in most cases, they are ignored.

According to Samsonova et al. (2017), the interpolation at some point is performed using only its subpopulation of available observations which makes the method independent of trends and then offers hope for a significant expansion of the application areas. Also, it is notable that the summary of the EBK algorithm is a heuristic algorithm at present.

This method is recent but has been employed in the mapping of the distribution of organic carbon in agricultural lands (Samsonova et al. 2017), for determination of radiation contamination levels after the Fukushima nuclear power plant (NPP) accident (Gribov and Krivoruchko 2012) and also for benthos mapping (Mulcan et al. 2015). However, there are demerits in using the method such as slow processing than other kriging methods when generating the interpolated raster. Also, processing time rapidly increases as the number of input points, the subset size or the overlap factor increase and log transformation are only sensitive to outliers.

Generally, the semivariogram model was employed to estimate the spatial autocorrelation of the prediction (Webster and Oliver 2007), and the equation expresses it,

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^n [Z(X_i) - Z(X_i + h)]^2 \quad (2)$$

where  $\gamma(h)$  is the semi-variance,  $N(h)$  is the point group number at distance  $h$ ,  $Z(x_i)$  is the numerical value at position  $x_i$  and  $Z(x_i + h)$  is the numerical value at a distance  $(x_i + h)$ .

## Second step

**Multivariate statistics** The two multivariate statistics adopted in this study are PCA and MLR.

**Principal component analysis** The PCA allows the grouping of similar variables into dimensions or principal components, without differentiating independent and dependent variables (Borůvka et al. 2005). Before implementing these multivariate statistics, a simple Pearson's correlation analysis was carried out using R software to check the correlations among soil properties and the selected soil-environmental covariates. For the PCA, only principal components (PCs), factors with eigenvalues  $> 1$  were considered to be contributing in explaining to the variability in the soil properties. In the present study, in the closest possible way, the research evaluated variables that should express in the details the variability observed between soil properties and soil-environmental covariates, revealing how significant soil-environmental covariates are in predicting soil property distribution in a given geographical position (Malinowski

2002; Shukla et al. 2006). In PCA, variables with high factor loading (Eigen  $> 1$ ) can be used to suggest the relationship between variables under each factor (Zeraatpisheh et al. 2019).

**Multiple linear regression** MLR technique uses several explanatory variables to predict the outcome of a response variable. The main reason behind this model is that the model tries to explain the spatial distribution of a dependent variable through a linear relationship between the explanatory variables (soil-environmental covariates) and the response variable (soil property) (Eq. 3).

$$y = a + \sum_{i=1}^n b_i * x_i \pm \epsilon_i \quad (3)$$

where for  $n$  number of predictors,  $y$  dependent variable (soil properties),  $x_i$  explanatory variables, independent variables or predictors (soil-environmental covariates),  $a$  intercept (constant term),  $b_i$  partial regression coefficients and  $\epsilon_i$  the model's error term (also known as the residuals).

This regression equation is used to model the spatial distribution of the variable of interest-based on the independent variables. Also, one soil property was modelled at a time as the response (dependent) variable with the generated matrix values of the soil-environment covariates from SAGA-GIS. For each model in MLR, the coefficient of determination ( $R$ -squared) is used to explain the extent of variation in the soil-environmental covariates. In this study, all the soil-environmental covariates were applied to each targeted soil property, and the predictor(s) that were significant at a 5% and 10% significance level were noted. The "lm" function implemented in the R software was used in the MLR analysis.

## Final step

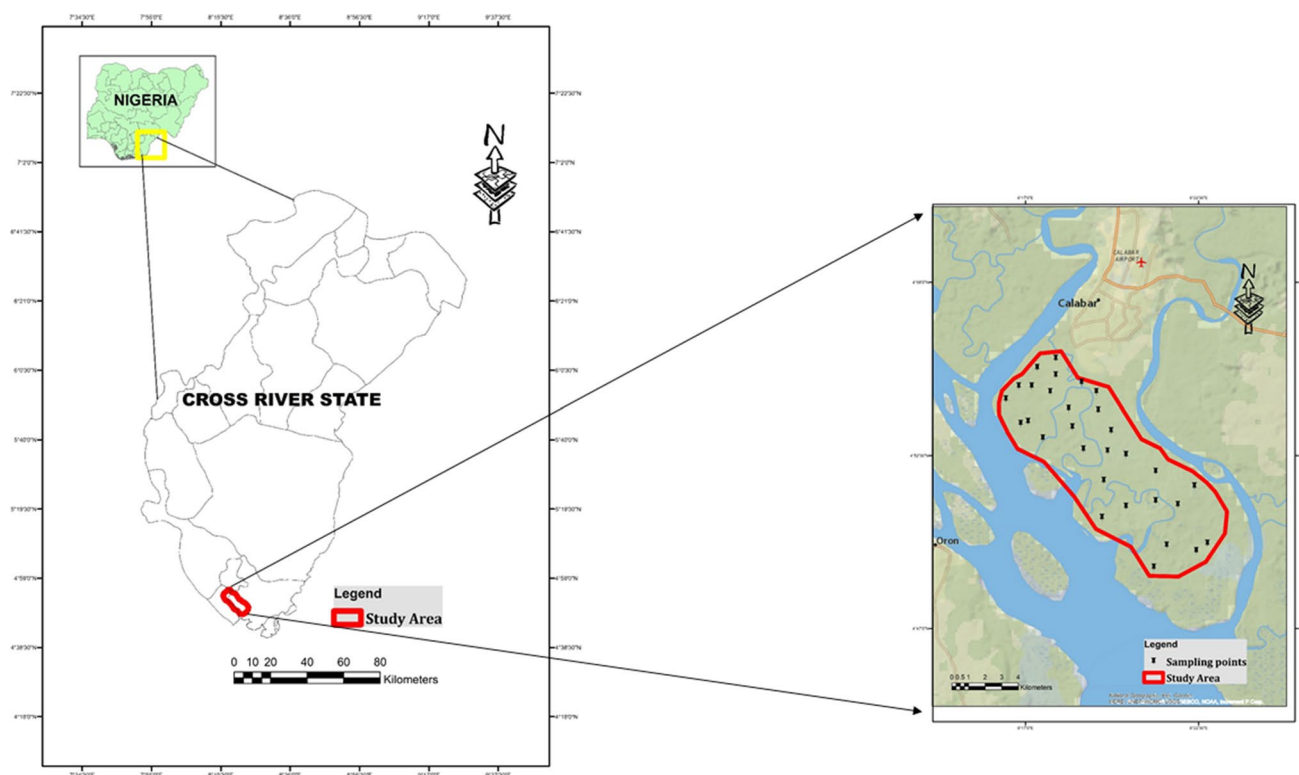
The EBK interpolation method was compared with MLR using the coefficient of determination ( $R$ -squared) to evaluate the model with the best fit for each targeted soil property.

## Case study

### The study area

The present research was carried out in Calabar, Cross River State, in the southeastern region of Nigeria. This area extends from latitude  $4^\circ 51' 45.86''$  N and longitude  $8^\circ 19' 50.69''$  E (Fig. 1) and spreads over an area of approximately  $200 \text{ km}^2$  with an elevation of about 44 m above sea level (John et al. 2018). It is under the humid tropical rainforest zones, marked with two distinct seasons (rainy and dry seasons). The area receives the mean annual rainfall of above 3500 mm with a temperature range between 22 and  $30^\circ \text{C}$ ,





**Fig. 1** Map of the study area showing sampling points

and relative humidity of 83% (John et al. 2018). The main land-uses include rain-fed cultivation of tree crops and arable crops. The predominant crops are oil palm trees, banana, maize, sugar cane, cassava, groundnut and vegetable crops. The central landscape units in the study area are relatively flat terrain.

The soils of the study area are developed from a coastal plain sand parent material (Akpan-Idiok 2012; John et al. 2018; Afu et al. 2019). They are characterized by udic moisture regime and isohyperthermic temperature regimes, respectively (Soil Survey Staff 2014). In the area, to a great extent, unconsolidated materials occurred with high sand and silt consistency (Akpan-Idiok 2012; John et al. 2018; Afu et al. 2019). The coarse-texture and the low activity clays of the distribution of the soil places the soil order into Inceptisols and Ultisols, respectively (Esu 2005).

### Soil sampling

In the present study, the sampling regime was done in the year 2018. The total of thirty ( $n=30$ ) composite samples was collected through stratified random sampling at a depth of 0–30 cm with the aid of a soil auger and a hand-held global positioning system (GPS), packaged into a Ziploc bag and transported to the laboratory for analysis.

### Laboratory analysis

For laboratory analysis, the soil samples were air-dried and sieved through a 2-mm sieve before being analyzed. For particle size analysis, the hydrometer method described by Bouyoucos (1962) was used. Soil pH in water was performed by the method outlined by Udo et al. (2009) at the ratio of 1 g of soil to 2.5 ml of water mixture. Organic carbon, available phosphorus, total nitrogen, exchangeable bases ( $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$  and  $\text{K}^{+}$ ), exchangeable acidity and base saturation were determined by the methods outlined by Okalebo et al. (2002) and Udo et al. (2009).

### Environmental covariates

To model soil properties variation, nine (9) different sets of soil-environmental covariates were derived (Table 1; Fig. 2a, b). These covariates were derived from the digital elevation model (DEM) obtained at the spatial resolution of 30 m from Shuttle Radar Topography Mission (SRTM) data (U.S Geological Survey 2020) and processed using SAGA-GIS free software terrain analysis toolbox. The derivatives are elevation (El), slope (Slp), aspect (As), analytical hillshading (Ah), plan curvature (PICur), profile curvature (PrCur), topographic wetness index (TWI), convergence index (CI) and LS-factor (Ls) were derived

**Table 1** Soil-environmental covariates used in the study

Data source	Soil-environmental covariates	Type	Significance
Terrain	Elevation	El	Climate, vegetation and energy potential
	Slope	Slp	Surface and subsurface flows, flow speed and erosion rate, precipitation, vegetation, geomorphology, soil water content and land use capacity
	Profile curvature	PrCur	Profile curvature is the rate of change of slope in a downslope direction. It characterizes changes in flow acceleration that may differentiate erosion and deposition zones in landscapes
	Plan curvature	PlCur	Convergent/divergent flows, soil water content, soil characteristics, flow acceleration, erosion rate/deposition and geomorphology
	Aspect	As	Solar radiation, evapotranspiration, flora and fauna distribution and abundance
	LS-factor	Ls	Surface flow volume
	Topographic wetness index	TWI	A measure of the topographic control on soil wetness
	Analytical hillshading	Ah	Hillshading is a technique for visualizing terrain determined by a light source and the slope and aspect of the elevation surface
	Convergence index	CI	Is used to distinguish flow convergent areas from divergent ones in the DTM at initial spatial resolution

using SAGA-GIS software (Olaya 2004). These soil-environmental covariates were selected due to their proven correlation with soil properties (Bishop and McBratney 2001; Penížek and Boruvka 2006).

### Cross-validation

In the cross-validation of the predicted map, 50% of the data to estimate the trend and autocorrelation of the models were used. The subsetting of the data was carried out in the geostatistic tool option in the Arcgis 10.7 software (Gribov and Krivoruchko 2012). The model was then evaluated using the coefficient of determination ( $R^2$ ) and root-mean-square error (RMSE). The  $R^2$  was estimated by plotting the predicted against the observed, while the RMSE was automatically calculated from EBK interpolation method and they are both expressed in these equations,

$$R^2 = 1 - \frac{\sum_{i=1}^n (pi - oi)^2}{\sum_{i=1}^n (pi - oi)^2} \quad (4)$$

where  $pi$  = predicted values,  $oi$  = observed values

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (pi - oi)^2} \quad (5)$$

Interpretatively, a good model fit for  $R^2$  is equal or close to 1 while for RMSE, close to 0. Furthermore, Li et al. (2016) proposed a classification criterion for  $R^2$  values:  $R^2 < 0.50$  (unacceptable prediction),  $0.50 < R^2 < 0.75$  (acceptable prediction) and  $R^2 > 0.75$  (good prediction). This classification was considered for this study.

### Statistical analysis

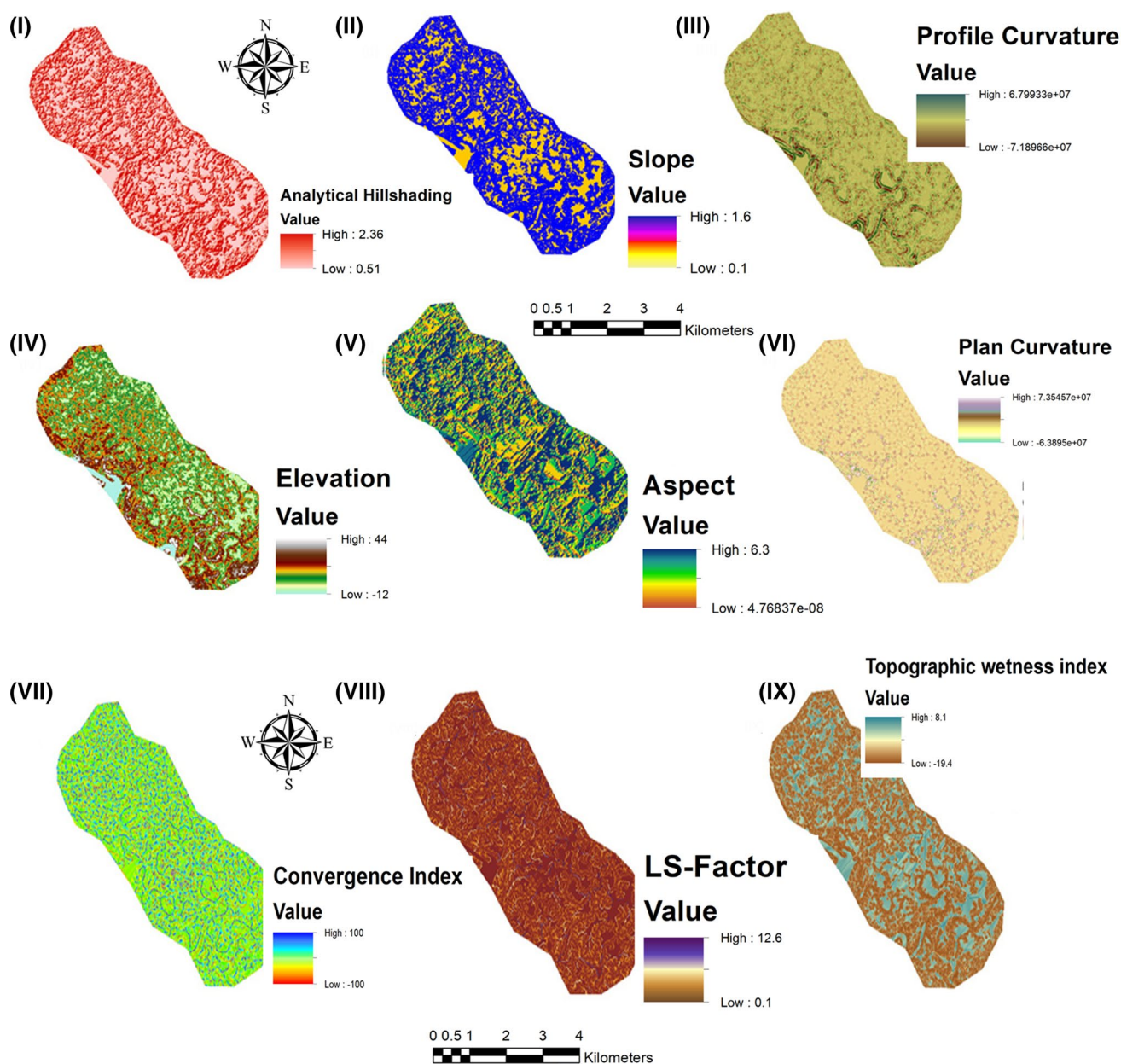
Descriptive statistics (mean, minimum, maximum, standard deviation, skewness and kurtosis) of soil properties and soil-environmental covariates were determined. Pearson correlation, PCA and MLR were determined to identify the soil property distribution and relationships between soil properties and soil-environmental factors. All statistical analysis was done using the R software (R Core Team 2019; Version 4.0).

## Results and discussion

### Descriptive statistics

Descriptive statistics of the soil properties and soil-environmental covariates are presented in Table 2. The sand content ranged from 60 to 83%; silt content ranged from 17 to 27%; and clay content ranged from 2.7 to 29%. The soil pH ranged from 5.5 to 5.9, with a mean of 5.5. While soil organic carbon content ranged from 0.6 to 3.1%. The frequency distributions of the soil properties showed that particle size distribution (sand, silt and clay) and organic carbon are positively skewed, while pH is negatively skewed. The result presented a pH with the highest skewness and kurtosis among all soil properties. The result obtained in the study is similar to the report by John et al. (2019). The variability obtained in all the soil properties may be attributed to the soil's inherent property associated with the parent material from which soils are developed.

Among all soil properties, sand, silt and clay contents gave a high standard deviation, which indicates a wide range of distribution values across the study area. The average sand



**Fig. 2** Soil-environmental covariates derived from digital elevation model at 30 m spatial resolution

content of 70.8% suggests that the soils of the study area are coarse-textured as they possess sand content  $\geq 70\%$ . The result corroborates with the report by Akpan-Idiok (2012), Akpan et al. (2017) and Afu et al. (2019). Furthermore, this type of soil cannot retain water and essential plant nutrients (Akpan-Idiok 2012). The soil is acidic with a pH of 5.5, and this may have been contributed by high annual rainfall above 3500 mm and excessive usage of NPK fertilizers (Wallace 1994; Wei et al. 2020). More so, studies have revealed that temperature and precipitation are essential factors that control soil pH (Cheng-Jim et al. 2014; Zhang et al. 2019; Dharumarajan et al. 2017). Other soil properties such

as OC, TN and the basic cations ( $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$  and  $\text{K}^{+}$ ) were all low when compared to their standard rating (Adaikwu and Ali 2013). While P content was moderate in the soil as it ranges between 8 and 20 mg/kg. This because in the topsoil, the P content is always high and most P fertilizer applied by the farmers is insoluble. The exchangeable acidity was low ( $< 2.52$  cmol/kg) and effective cation exchange capacity value was moderate (range  $> 5.52$  cmol/kg) compared to rating established for productive soils (Adaikwu and Ali 2013). Furthermore, the mean of base saturation = 76.2% may refer to the fact that essential nutrients must have prevailed in available forms in the soil solution despite the low cation

**Table 2** Descriptive statistics of soil properties and soil-environmental covariates

	Mean	Minimum	Maximum	SD	Skewness	Kurtosis
Sand (%)	70.8	60	83	7.1	0.3	− 1.1
Silt (%)	17	10	27	4.6	0.3	− 0.9
Clay (%)	12.3	2.7	29	8.6	0.8	− 0.7
pH	5.5	4.5	5.9	0.3	− 1.8	5.5
OC (%)	1.8	0.6	3.1	0.8	0.3	− 1.5
TN (%)	0.2	0.1	0.3	0.1	0.2	− 1.4
P (mg/kg)	11	4.7	23.3	4.2	1.4	2.5
Ca <sup>2+</sup> (cmol/kg)	4.5	0.6	12.8	3.0	1.2	1.3
Mg <sup>2+</sup> (cmol/kg)	1.4	0.2	4	1.0	1.1	0.8
K <sup>+</sup> (cmol/kg)	0.1	0.1	0.2	0.1	0.5	− 0.5
Na <sup>+</sup> (cmol/kg)	0.1	0.1	0.2	0.1	0.4	− 0.9
Exch. acidity (cmol/kg)	1.7	0.3	5.2	1.5	1.1	− 0.1
ECEC (cmol/kg)	7.6	4.7	14.6	2.3	1.6	2.7
BS (cmol/kg)	76.2	27.3	96	22.1	− 0.9	− 0.9
El (m)	7.1	3.2	10.2	1.9	− 0.34	− 0.87
Ah (deg)	1.4	0.8	2.3	0.5	0.5	− 1.2
S (deg)	1.1	0.1	1.6	0.6	− 0.6	− 1.4
As (rad)	2.9	1.0	5.5	1.5	0.6	− 1.4
PICurv	592,522.5	− 9,228,767	6,253,365	3,680,354.8	− 0.7	0.3
PrCur	− 342,838.2	− 9,365,311.7	5,524,231.4	3,610,322.3	− 0.6	0.4
CI	2.3	− 33.3	40.1	15.8	− 0.1	0.5
TWI	− 10.2	− 17.2	1.4	6.4	0.5	− 1.5
Ls	1.6	0.1	4.4	1.0	0.5	1.6

OC organic carbon, Mg magnesium, deg degrees, TN total nitrogen, K potassium, m metres, P phosphorus, Na sodium, rad radians, Ca calcium, exch. acidity active acidity, ECEC effective cation exchange capacity and BS base saturation

reserves (Akpan-Idiok 2012). This present result may also be attributed to the discriminate use of soil input materials (e.g. fertilizers and herbicides) for crop production by the active land users in the region and resulting to leaching of these basic cations (Sharu et al. 2013). The results obtained here corroborate with other works in the region on a similar type of soils (Akpan-Idiok 2012; John et al. 2018; Afu et al. 2019; Akpan et al. 2017). The study further explains that the farmers intensively use the soils in the regions without good soil management programme.

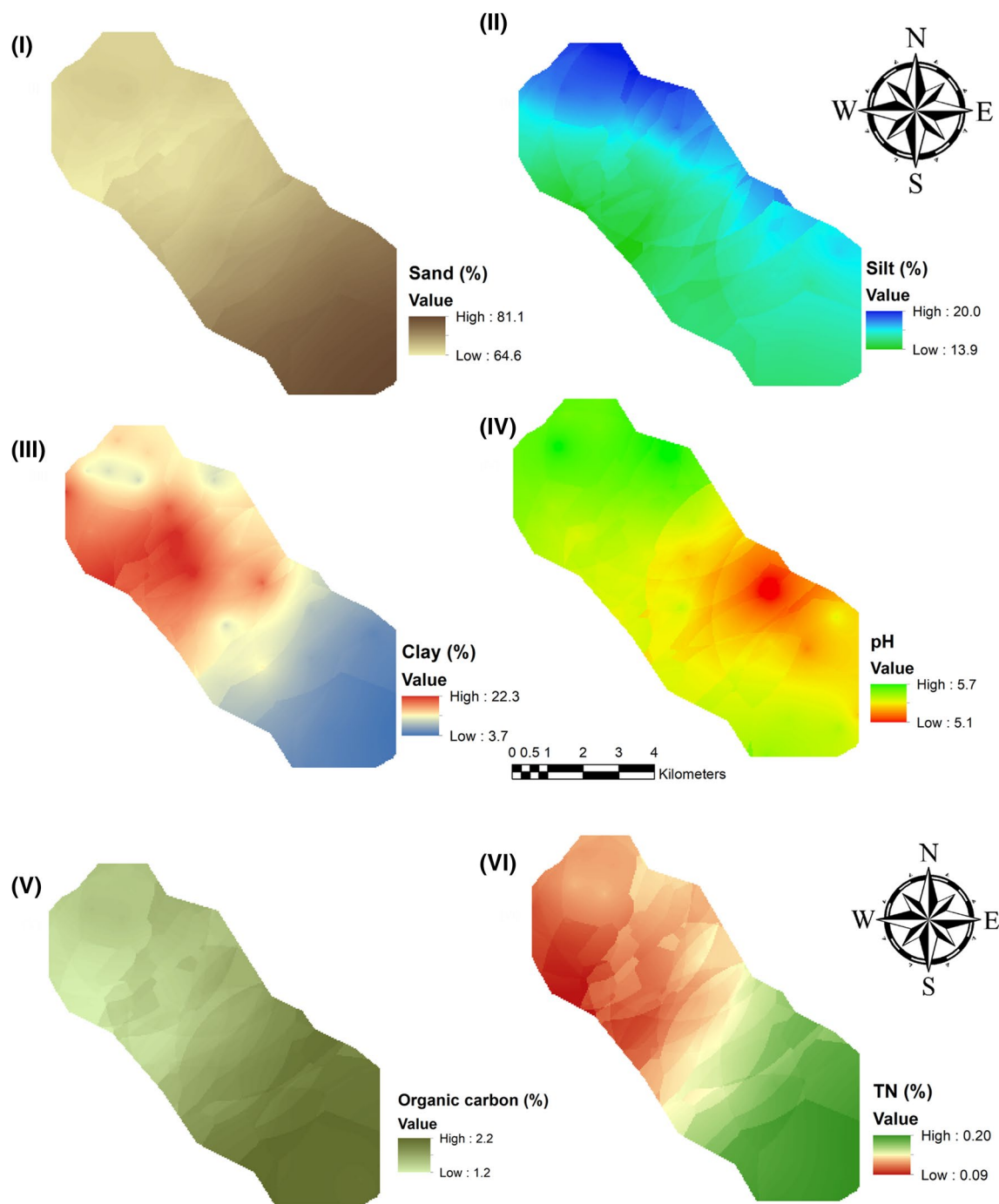
### Empirical Bayesian kriging

The EBK interpolated soil property maps are presented in (Fig. 3a–c). The map revealed that the highest sand content was observed in the south direction, while the lowest sand fraction was observed at the north angle and down towards the central region (Fig. 3a, I). The silt fraction was highest in the northern direction and the northern-west direction for clay (Fig. 3a, II–III). The result obtained here may be attributed to gradual transportation and deposition of sediments (Akpan-Idiok 2012).

The highest pH value in the study area was observed in the east direction (Fig. 3a, IV). Furthermore, the highest organic carbon contents were predominantly seen in the southern direction (Fig. 3b, V). The result obtained here followed a similar pattern with that of Zeraatpisheh et al. (2019). They reported high organic carbon from the centre to the southwestern direction in the semi-arid region of Iran using inverse distance weighting interpolation. Other soil properties maps revealed their respective distribution over the study area (Fig. 3b, c).

### Cross-validation

The results of the cross-validation are presented in Table 3. With the use of EBK maps, the study affirmed that the interpolation technique was best to understand the spatial distribution of the soil properties in the studied area (Li and Heap 2011). The interpolation method that yielded the highest  $R^2$  with the corresponding lowest RMSE values are Mg ( $R^2=0.778$ , RMSE=0.866), K ( $R^2=0.637$ , RMSE=0.017), P ( $R^2=0.629$ , RMSE=0.06), pH ( $R^2=0.675$ , RMSE=0.267) and TN ( $R^2=0.721$ , RMSE=0.755). The good model fit obtained in the study



**Fig. 3** Empirical interpolated soil property maps

explained how the EBK interpolation method analyzed different uncertainties that are influenced by several factors such as edaphic activities, various soil management methods, fertilizer application rates, tillage systems and others (Samsonova et al. 2017). This was the case in the present study site. The good outputs observed in EBK have

confirmed the works of Adhikary et al. (2011), Fabijańczyk et al. (2017), Beguin et al. (2017) and Yan et al. (2019), respectively on EBK being more effective compared to other kriging methods. The result also revealed that EBK interpolation model could effectively predict soil properties variation in a relatively flat terrain condition.



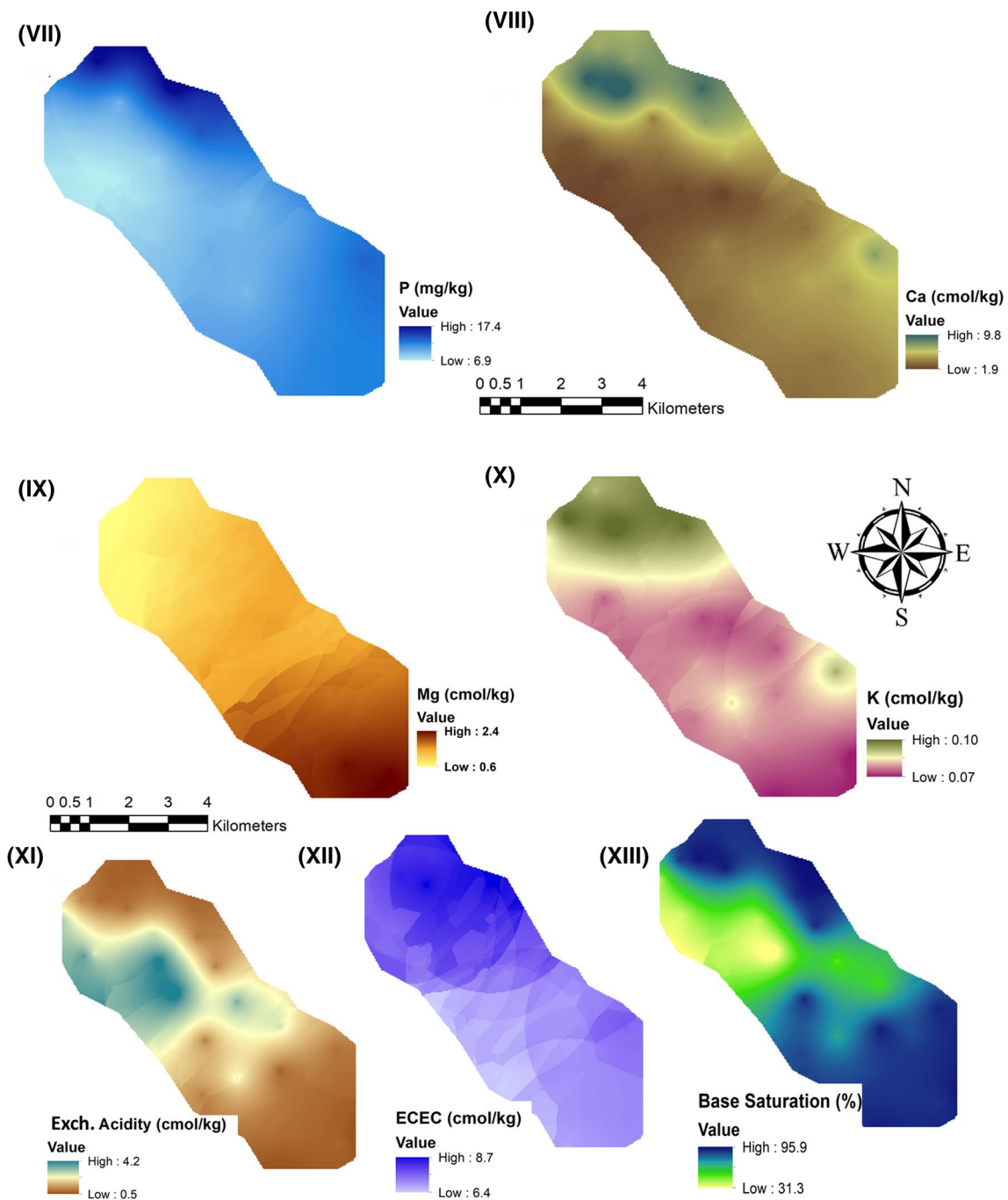


Fig. 3 (continued)

**Correlation matrix**

Presented in Fig. 4 is the correlation matrix plot between soil properties and soil-environmental covariates. Sand showed a weak negative correlation with plan curvature, profile curvature and convergence index. Silt presented a positive but significant correlation with Ah, SIp and PICur

but negatively correlated with As and TWI. The result obtained here is in contrast with the report obtained by Kokulan et al. (2018) who reported a strong negative correlation between soil texture (sand, silt and clay) and El, PICur and PrCur in a gently undulating topography but in support with the works by Mosleh et al. (2016) on low relief condition. OC, TN, P, Ca<sup>2+</sup>, Mg<sup>2+</sup>, K<sup>+</sup>, exch. acidity

**Table 3** Cross-validation criteria for empirical Bayesian kriging (EBK) interpolation method among soil properties

Soil properties	R <sup>2</sup>	RMSE
Sand (%)	0.797	5.270
Silt (%)	0.573	4.410
Clay (%)	0.533	7.400
pH	0.675	0.267
Total nitrogen (%)	0.721	0.755
Phosphorus (mg/kg)	0.629	0.060
Organic carbon (%)	0.604	3.497
Calcium (cmol/kg)	0.740	2.520
Magnesium (cmol/kg)	0.778	0.866
Potassium (cmol/kg)	0.637	0.017
Exchangeable acidity (cmol/kg)	0.696	1.260
Effective cation exchange capacity (cmol/kg)	0.549	2.310
Base saturation (%)	0.626	16.40

R<sup>2</sup>, coefficient of determination; RMSE, root-mean-square error

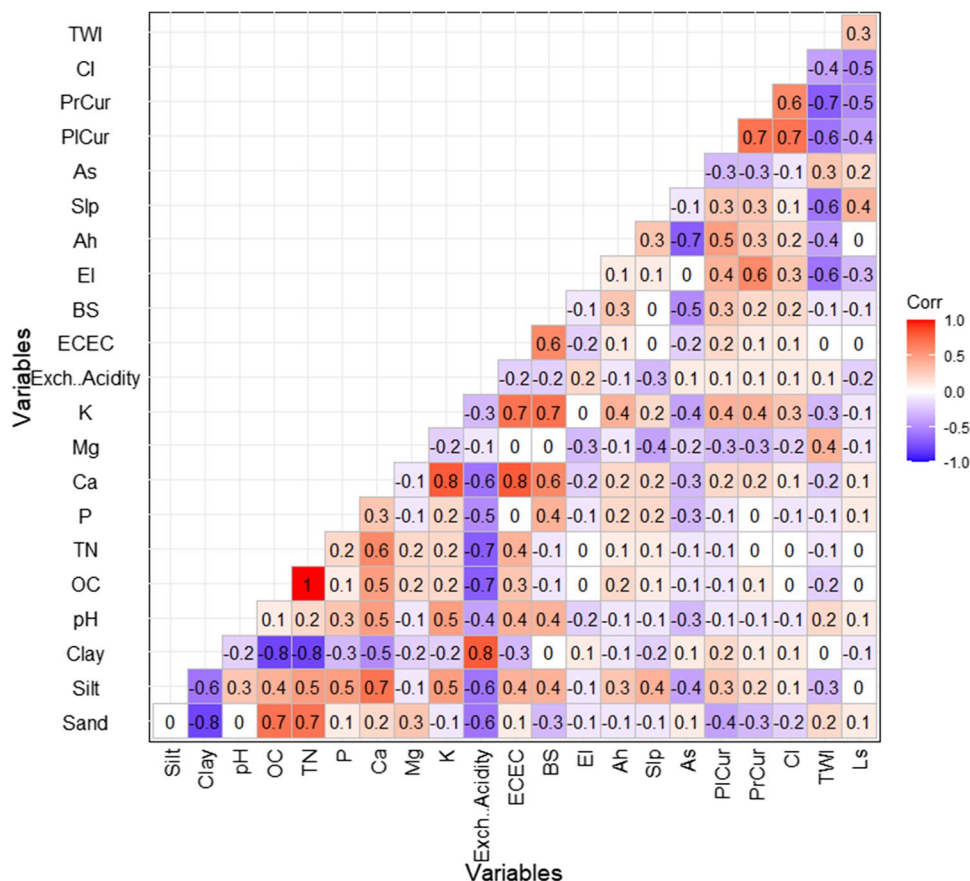
and ECEC all weak to no significant correlation with the soil-environmental covariates. Magnesium showed a negative but significant correlation with all the nine environmental attributes at a significant level of 5%. This is in contrast with the study conducted by Kokulan et al. (2018) on a heterogeneous landscape position.

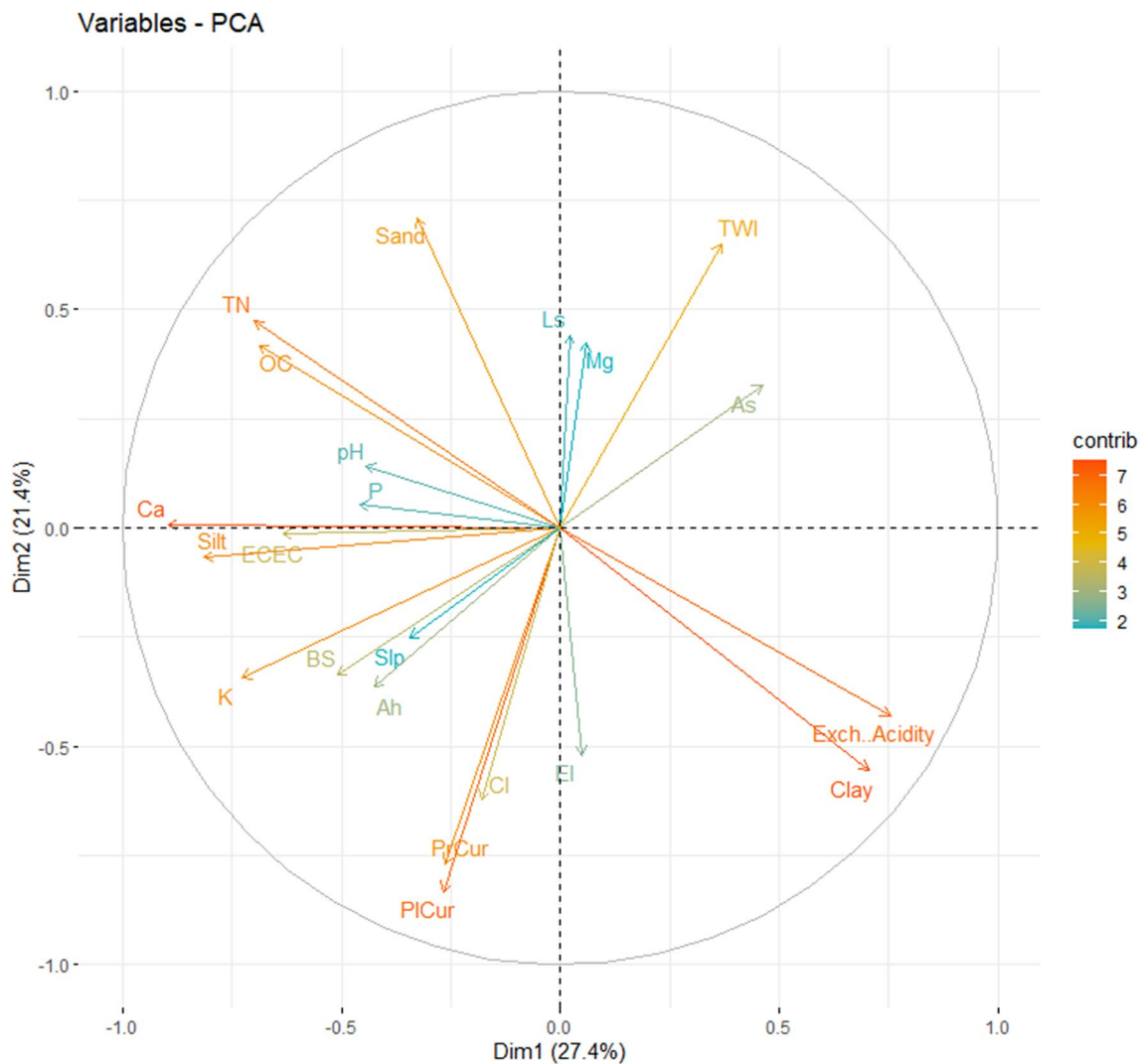
**Principal component analysis**

In Fig. 5 and, the factor analysis is also known as PCA, presents five important PCs to interpret the variability in the observations. The analysis showed that PC1 and PC2 could explain 52.6% of the total variance (Figs. 5, 6). Then, when adding the other four PCs, a total of 78.1% is described. Figure 6 presented the obtained PCs, but only five PCs were selected.

Principal component (PC1) explained 30.1% of the total variation in the observations, and it is majorly contributed by all the soil properties (sand, silt, clay, pH, OC, TN, Ca<sup>2+</sup>, K<sup>+</sup>, exch. acidity, ECEC and BS). PC2 explained 22.5%

**Fig. 4** Correlation matrix plot ( $p < 0.05$ ) ( $n = 30$ ), significance level of  $\alpha = 0.05$





**Fig. 5** Principal component analysis of the variables

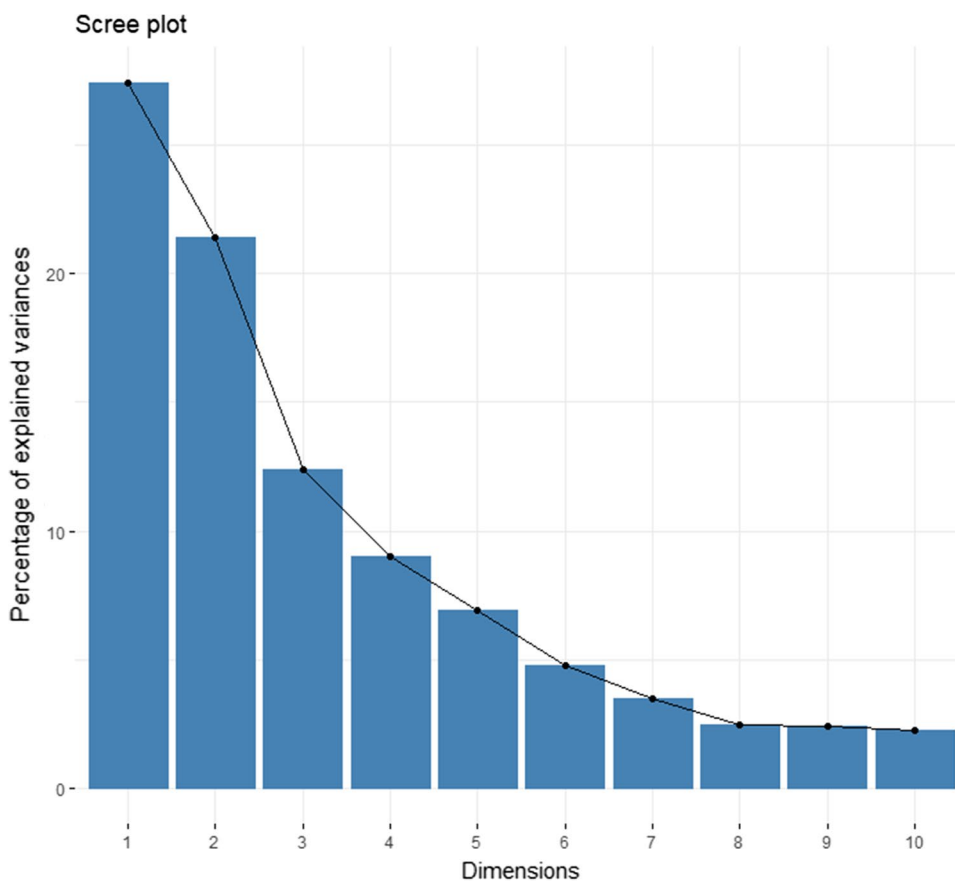
of the total variance, with significant contributions from soil-environmental covariates (El, Ah, PICur, PrCur, CI and TWI) and sand content. PC3 explains about 10.5% of the total variability in the observations, being highly contributed to by the clay content and elevation. PC4 described approximately 8.8% of the total variability, which is contributed by Slp and Ls. Furthermore, PC5 yielded approximately 6.2% of the overall variability, with major influence from As.

Soil properties that dominated PC1 explained the high correlation observed between soil properties (Fig. 4). Similarly, PC2 was seen to be dominated by soil-environmental covariates. The weak to no correlation between some soil properties and soil-environmental covariates obtained in

the correlation matrix plot (Fig. 4) was reconfirmed in the PCA (Fig. 5). This result may be attributed to lack of influence of the flat terrain (e.g. slope gradient = 0.1–1.6) to contribute on the spatial variability of soils in the region. More so, in a relatively stable topography, soil properties are actively influenced by their properties rather than the topographic condition. The result obtained here is in harmony with the works done by Zeraatpisheh et al. (2019) and Mosleh et al. (2016) done elsewhere. The present research, therefore, inferred that additional soil-environmental covariates such as geology, land use, precipitation and evapotranspiration variables being introduced could improve the modelling soil properties in low relief conditions.



**Fig. 6** A scree plot showing the percentage of variances explained by each principal component



### Multiple linear regression (MLR)

The result of the MLR is presented in Table 4. All the soil-environmental covariates were applied to each soil property one at a time. Thirteen models were obtained. An acceptable prediction ( $0.50 < R^2 < 0.75$ ) was observed for  $Mg^{2+}$ , while for other soil properties yielded unacceptable prediction ( $R^2 < 0.5$ ) was observed. The output observed for  $Mg^{2+}$  may be attributed to its high mobility through the action of mass flow and water imbalance (Gransee and Fühns 2013). On the other hands, the low  $R^2$  value answered the hypothetical questions, i.e. the expectation that little or no influence of terrain attributes on soil properties distribution in relatively flat topography. This was true for the other soil properties except for  $Mg^{2+}$ . The results obtained here corroborate with the poor performance of MLR in similar study elsewhere (Mosleh et al. 2016).

The results in this present study demonstrated that the coastal plain parent material reflects on the properties of the soil and consequently, a good indicator of soil fertility status. Furthermore, in a flat terrain system, soil-environmental covariates contribute little or no variability in the

soil properties distribution. However, the results obtained in the PCA are in corroboration with those of Khaledian et al. (2017), who used PCA to discriminate soil indicators taking into account spatial variables and that of Zeraatpisheh et al. (2019) who combined PCA and inverse distance weighting interpolation in predicting soil properties in the dry arid region of Iran. Also, the result of the PCA corresponds with the work of Souza et al. (2018) who also used it in dataset reduction and for soil properties modeling in Ultisols.

### Comparison between empirical Bayesian kriging (EBK) and multiple linear regression (MLR)

In comparing the best models between empirical Bayesian kriging and the multiple linear regression for estimating the soil properties distribution, the study adopted the Li et al. (2016) classification criterion for  $R^2$  (Table 5). The basis for the comparison between EBK and MLR was to answer the hypothetical questions that low relief condition contributes little or no amount in soil properties variable and that the intrinsic properties of the soils better explain itself. Besides

**Table 4** Best multiple linear equations between soil parameters and soil-environmental covariates

Soil properties	Regression equation	Soil-environmental variable of importance	R <sup>2</sup>
Sand	$51.3 + 0.5 \times EI + 0.6 \times Ah + 20.2 \times SIp + 0.94 \times As - 2.3 \times 10^{-6} \times PICur - 2.55 \times 10^{-8} \times PrCur + 0.09 \times CI + 0.95 \times TWI - 6.6 \times Ls$	PICur*	0.305
Silt	$10.2 - 0.46 \times EI - 1.47 \times Ah + 5.97 \times SIp - 1.16 \times As + 5.08 \times 10^{-7} \times PICur - 2.7 \times 10^{-7} \times PrCur - 0.0092 \times CI - 0.397 \times TWI + 0.62 \times Ls$	Slp*	0.315
Clay	$38.5 - 0.0362 \times EI - 5.25 \times Ah - 2.62 \times SIp + 0.22 \times As + 1.8 \times 10^{-7} \times PICur + 310^{-6} \times PrCur - 0.081 \times CI - 0.55 \times TWI + 5.96 \times Ls$		0.180
pH	$6.10 + 0.0015 \times EI - 0.236 \times Ah + 0.063 \times SIp - 0.125 \times As + 6.67 \times 10^{-8} \times PICur - 1.63 \times 10^{-9} \times PrCur + 0.014 \times CI + 0.014 \times TWI + 0.095 \times Ls$	Ah*, As*	0.219
OC	$2.1 - 0.057 \times EI + 0.72 \times Ah + 0.64 \times SIp - 0.071 \times As - 1.58 \times 10^{-7} \times PICur + 4.2 \times 10^{-8} \times PrCur + 0.0030 \times CI - 0.028 \times TWI + 0.031 \times Ls$		0.154
TN	$0.0025 - 0.0025 \times EI + 0.048 \times Ah + 0.056 \times SIp + 0.0049 \times As - 1.2 \times 10^{-7} \times PICur - 7.1 \times 10^{-9} \times PrCur + 0.00053 \times CI - 0.0032 \times TWI - 0.024 \times Ls$		0.126
P	$8.2 - 0.22 \times EI - 0.058 \times Ah + 1.66 \times SIp - 1.21 \times As - 3.75 \times 10^{-7} \times PICur + 5.51 \times 10^{-7} \times PrCur - 0.0061 \times CI - 0.258 \times TWI + 1.03 \times Ls$	As	0.211
Ca	$8.2 - 0.782 \times EI - 0.461 \times Ah - 7.75 \times SIp - 0.37 \times As + 1.8 \times 10^{-7} \times PICur + 2.3 \times 10^{-7} \times PrCur + 0.017 \times CI - 0.583 \times TWI + 3.43 \times Ls$	TWI*	0.290
Mg	<b><math>2.2 - 0.0856 \times EI - 0.00595 \times Ah + 3.47 \times SIp - 0.37 \times As + 1.8 \times 10^{-7} \times PICur + 2.3 \times 10^{-7} \times PrCur + 0.017 \times CI - 0.583 \times TWI + 3.43 \times Ls</math></b>	<b>Slp*, As*, PICur</b>	<b>0.509</b>
K	$0.12 - 0.0035 \times EI - 0.00091 \times Ah - 0.029 \times SIp - 0.0042 \times As + 8 \times 10^{-10} \times PICur + 2.1 \times 10^{-10} \times PrCur + 0.00021 \times CI - 0.00184 \times TWI - 0.0128 \times Ls$	PrCur*	0.359
Exch. acidity	$6.34 - 0.070 \times EI - 0.0437 \times Ah - 3.02 \times SIp - 0.153 \times As + 2.72 \times 10^{-8} \times PICur + 3.16 \times 10^{-9} \times PrCur - 0.00241 \times CI + 0.034 \times TWI - 0.031 \times Ls$	Slp*	0.217
ECEC	$16.7 - 0.624 \times EI - 0.0826 \times Ah - 7.22 \times SIp - 0.249 \times As + 2.8 \times 10^{-7} \times PICur + 1.73 \times 10^{-7} \times PrCur - 0.00179 \times CI - 0.30 \times TWI - 1.88 \times Ls$	PICur, EI	0.220
BS	$46.8 - 2.09 \times EI + 4.11 \times Ah + 21.4 \times SIp - 3.12 \times As - 2.2 \times 10^{-6} \times PICur - 6 \times 10^{-8} \times PrCur - 0.00179 \times CI - 0.30 \times TWI - 1.88 \times Ls$	EI, As	0.153

Bold R<sup>2</sup> gave a good model fit; *p* < 0.1; \**p* < 0.05; \*\**p* < 0.01, \*\*\**p* < 0.001

**Table 5** Comparison the best model between EBK and MLR

Soil properties	R <sup>2</sup>	
	EBK	MLR
Sand (%)	0.797	0.305
Silt (%)	0.573	0.315
Clay (%)	0.533	0.180
pH	0.675	0.219
Total nitrogen (%)	0.721	0.154
Phosphorus (mg/kg)	0.629	0.126
Organic carbon (%)	0.604	0.211
Calcium (cmol/kg)	0.740	0.290
Magnesium (cmol/kg)	0.778	0.509
Potassium (cmol/kg)	0.637	0.359
Active acidity (cmol/kg)	0.696	0.217
Effective cation exchange capacity (cmol/kg)	0.549	0.220
Base saturation (%)	0.626	0.153

that, the study did not consider modelling with variables that have only a significant correlation with soil-environmental variables; preferably, all the nine soil-environmental variables were applied as a response variable to predict each targeted soil property. The reason is that according to the PCA, PC1–PC5, the soil-environmental variables made significant contributions to their PCs hence the reason to engage them in the modelling regime. The low R<sup>2</sup> value obtained in MLR answered the hypothetical questions. The poor performance of MLR in the study is similar to that of Mosleh et al. (2016) in low relief condition. Furthermore, the study revealing that EBK performed better than MLR only explained geological factors which are an intrinsic mechanism explains more of soil properties variation than terrain derivatives in the region.

Conversely, MLR gave a an acceptable prediction for only Mg<sup>2+</sup> ( $2.2 - 0.0856 \times EI - 0.00595 \times Ah + 3.47 \times SIp - 0.37$

$\times \text{As} + 1.8 \times 10^{-7} \times \text{PICur} + 2.3 \times 10^{-7} \times \text{PrCur} + 0.017 \times \text{CI} - 0.583 \times \text{TWI} + 3.43 \times \text{Ls}$ ,  $R^2 = 0.509$ ) with SIp, As and PICur as the only soil-environmental factors as the only relatively important variables. These may be attributed to the dynamic ecological environment, contrasting land use and low relief condition.

## Conclusion

In conclusion, this research aims to evaluate the capability of EBK using multivariate statistical analysis, in combination with soil-environmental covariates, to identify the spatial distribution of soil properties. General spatial patterns of soil properties (sand, silt, clay, pH, TN, OC, P,  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ , exch. acidity, ECEC and BS) can be predicted easily by using digital soil mapping techniques. The soils under study are coarse-textured with low natural fertility status. The EBK interpolation presented acceptable and good predictions for all the soil properties. However,  $\text{Mg}^{2+}$ ,  $\text{K}^+$ , P, pH and TN with lower RMSE, closer to 0 is said to be suitably predicted by EBK technique.

The PCA results showed that PC1 to PC5 could describe approximately 78.1% of the total variability in the soil properties. The result corresponds to the correlation matrix result. Furthermore, in MLR, each targeted property was predicted using the nine selected soil-environmental covariates in a low relief condition. The result showed that 50.9% of the soil-environmental covariates could largely explain the variability in  $\text{Mg}^{2+}$ , while for other soil properties, an unacceptable model fit was obtained.

Consequently, the study concludes that  $\text{Mg}^{2+}$  can be the best-predicted soil property because it follows a definite spatial pattern in flat terrain conditioned by SIp, As and PICur. This finding is paramount because  $\text{Mg}^{2+}$  is an essential nutrient element required in the tree crops grown in the study location. More so, for other soil properties, EBK was best in estimating their distribution.

Recommendation for further research, it should be fascinating to test different interpolation approaches such as regression kriging, ordinary kriging, radial basis functions and other machine learning algorithms with more soil-environmental factor aside from the ones generated from DEM to note if other soil properties do or do not correspond with  $\text{Mg}^{2+}$  and other soil properties as the best soil property predictor using multivariate statistical analysis. Furthermore, to improve the accuracy of the estimations through MLR, more observation points are recommended to interpret better the performance of the spatial mapping technique over low-relief areas.

**Acknowledgements** Mr John Kingsley would like to acknowledge the PhD scholarship and internal Grant No. 21130/1312/3131 offered by the Department of Soil Science and Soil Protection, Czech University of Life Sciences, Prague (CZU).

## Compliance with ethical standards

**Conflict of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

- Adaikwu AO, Ali A (2013) Assessment of some soil quality in Benue State. *Niger J Soil Sci* 23:66–75
- Adhikary PP, Dash J, Bej R, Chandrasekaran H (2011) Indicator and probability kriging methods for delineating Cu, Fe, and Mn contamination in groundwater of Najafgarh Block, Delhi, India. *Environ Monit Assess* 176:663–676
- Afu SM, Isong IA, Awaogu CE (2019) Agricultural potentials of flood-plain soils with contrasting parent material in Cross River State, Nigeria. *Glob J Pure Appl Sci* 25:13–22
- Akpa SI, Odeh IO, Bishop TF, Hartemink AE (2014) Digital mapping of soil particle-size fractions for Nigeria. *Soil Sci Soc Am J* 78:1953–1966
- Akpan JF, Aki EE, Isong IA (2017) Comparative assessment of wetland and coastal plain soils in Calabar, Cross River State. *Glob J Agric Sci* 16:17–30
- Akpan-Idiok AU (2012) Physicochemical properties, degradation rate and vulnerability potential of soils formed on coastal plain sands in southeast, Nigeria. *Int J Agric Res* 7:358–366
- Beguería S, Spanu V, Navas A, Machín J, Angulo-Martínez M (2013) Modelling the spatial distribution of soil properties by generalized least squares regression: toward a general theory of spatial variates. *J Soil Water Conserv* 68:172–184
- Beguín J, Fuglstad GA, Mansuy N, Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: comparison of spatial and non-spatial statistical approaches. *Geoderma* 306:195–205
- Bishop TFA, McBratney AB (2001) A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma* 103:149–160
- Boettinger JL, Ramsey RD, Bodily JM, Cole NJ, Kienast-Brown S, Nield SJ, Saunders AM, Stum AK (2008) Landsat spectral data for digital soil mapping. In: Hartemink AE, McBratney A, Mendonça-Santos M (eds) *Digital soil mapping with limited data*. Springer, Dordrecht
- Borůvka L, Vacek O, Jehlička J (2005) Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. *Geoderma* 128:289–300
- Bouyoucos GJ (1962) Hydrometer method improved for making particle size analyses of soils. *Agron J* 54:464–465
- Cheng-Jim J, Yuan-He Y, Wen-Xuan H, Yan-Fang H, Smith J, Smith P (2014) Climatic and edaphic controls on soil pH in alpine grasslands on the Tibetan Plateau, China: a quantitative analysis. *Pedosphere* 24:39–44



- Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty, vol 497. Wiley, Hoboken
- de Carvalho Junior WD, Chagas CDS, Lagacherie P, Calderano Filho B, Bhering SB (2014) Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. *Rev Bras Ciência do Solo* 42:1–20
- Delarmelinda EA, de Souza Júnior VS, Wadt PGS, Deng Y, Campos MCC, Câmara ERG (2017) Soil-landscape relationship in a chronosequence of the middle Madeira River in southwestern Amazon, Brazil. *CATENA* 149:199–208
- Dharumarajan S, Hegde R, Singh SK (2017) Spatial prediction of major soil properties using random forest techniques—a case study in semi-arid tropics of South India. *Geoderma Reg* 10:154–162
- Esu IE (2005) Characterization, classification and management problems of the major orders in Nigeria. 26th Inaugural Lecture, Department of Soil Science University of Calabar, pp 38–59
- Esu IE, Uko U, Aki EE (2014) Morphological, physiochemical and mineralogical properties of soils developed from basalt at Ikom, cross river state, Nigeria. In: Proceeding of the 38th annual conference of soil science society of Nigeria, March 10–14, 2014. Uyo, Nigeria, pp 89–100
- Fabijańczyk P, Zawadzki J, Magiera T (2017) Magnetometric assessment of soil contamination in problematic area using empirical Bayesian and indicator kriging: a case study in Upper Silesia, Poland. *Geoderma* 308:69–77
- Gransee A, Führs H (2013) Magnesium mobility in soils as a challenge for soil and plant analysis, magnesium fertilization and root uptake under adverse growth conditions. *Plant Soil* 368:5–21
- Gribov A, Krivoruchko K (2012) New flexible non-parametric data transformation for trans-Gaussian kriging. In: Abrahamson P, Hauge R, Kolbjørnsen O (eds) *Geostatistics Oslo 2012*. Springer, Dordrecht, pp 51–65. [https://doi.org/10.1007/978-94-007-4153-9\\_5](https://doi.org/10.1007/978-94-007-4153-9_5)
- Gribov A, Krivoruchko K (2020) Empirical Bayesian kriging implementation and usage. *Sci Total Environ* 722:137290
- Hussain I, Shakeel M, Faisal M, Soomro ZA, Hussain M, Hussain T (2014) Distribution of total dissolved solids in drinking water by means of bayesian kriging and gaussian spatial predictive process. *Water Qual Expo Heal* 6:177–185
- Jenny H (1941) *Factors of soil formation*. McGraw-Hill, New York
- John K, Ayito EO, Odey S (2018) Interaction between some soil physicochemical properties and weather variables on sub-humid tropical rainforest soils of Cross River State, Southeastern Nigeria. *Annu Res Rev Biol* 29(6):1–12
- John K, Lawani SO, Esther AO, Ndiye KM, Sunday OJ, Penížek V (2019) Predictive mapping of soil properties for precision agriculture using geographic information system (GIS) based geostatistics models. *Mod Appl Sci* 13:60
- Khaledian Y, Kiani F, Ebrahimi S, Brevik EC, Aitkenhead-Peterson J (2017) Assessment and monitoring of soil degradation during land use change using multivariate analysis. *Land Degrad Dev* 28:128–141
- Kokulan V, Akinremi O, Moulin AP, Kumaramage D (2018) Importance of terrain attributes in relation to the spatial distribution of soil properties at the micro scale: a case study. *Can J Soil Sci* 98:292–305
- Krivoruchko K, Gribov A (2014) Pragmatic Bayesian kriging for non-stationary and moderately non-Gaussian data. In: Crisan D, Golden K, Holm DD, Lewis M, Nishiura Y, Tribbia J, Zubelli JP (eds) *Mathematics of planet earth*. Springer, Berlin, pp 61–64
- Li L, Lu J, Wang S, Ma Y, Wei Q, Li X, Cong R, Ren T (2016) Methods for estimating leaf nitrogen concentration of winter oilseed rape (*Brassica napus* L.) using in situ leaf spectroscopy. *Ind Crops Prod* 91:194–204
- Malinowski ER (2002) *Factor analysis in chemistry*. Wiley, New York, pp 1–432
- Mirzaei R, Sakizadeh M (2016) Comparison of interpolation methods for the estimation of groundwater contamination in Andimeshk-Shush Plain, Southwest of Iran. *Environ Sci Pollut Res* 23(3):2758–2769
- Mosleh Z, Salehi MH, Jafari A, Borujeni IE, Mehnatkesh A (2016) The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environ Monit Assess* 188:1–13
- Mulcan A, Mitsova D, Hindle T, Hanson H, Coley C (2015) Marine benthic habitats and seabed suitability mapping for potential ocean current energy siting offshore southeast Florida. *J Mar Sci Eng* 3:276–298
- Odeh IO, Crawford M, McBratney AB (2006) Digital mapping of soil attributes for regional and catchment modelling, using ancillary covariates, statistical and geostatistical techniques. *Dev Soil Sci* 31:437–454
- Okalebo JR, Gathua KW, Woomer PL (2002) *Laboratory methods of soil and plant analysis: a working manual*, 2nd edn. Sacred Africa, Nairobi, p 21
- Olaya V (2004) *A gentle introduction to SAGA GIS*. The SAGA User Group Press, Gottingen, pp 1–216
- Park SJ, Vlek LG (2002) Prediction of three-dimensional soil spatial variability: a comparison of three environmental correlation techniques. *Geoderma* 109:117–140
- Penížek V, Borůvka L (2006) Soil depth prediction supported by primary terrain attributes: a comparison of methods. *Plant Soil Environ* 52(9):424–430
- Petropoulos GP, Ireland G, Barrett B (2015) Surface soil moisture retrievals from remote sensing: current status, products & future trends. *Phys Chem Earth* 83–84:36–56
- Qin CZ, Zhu AX, Qiu WL, Lu YJ, Li BL, Pei T (2012) Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. *Geoderma* 171:64–74
- R Core Team. (2019). *R: a language and environment for statistical computing*. R
- Samsonova VP, Blagoveshchenskii YN, Meshalkina YL (2017) Use of empirical Bayesian kriging for revealing heterogeneities in the distribution of organic carbon on agricultural lands. *Eurasian Soil Sci* 50:305–311
- Sanchez PA (1977) Properties and management of soils in the tropics. *Soil Sci* 124(3):187
- Sharu MB, Yakubu M, Noma SS, Tsafe AI (2013) Characterization and classification of soils on an agricultural landscape in Dingyadi District, Sokoto State, Nigeria. *Niger J Basic Appl Sci* 21:137–147
- Shukla MK, Lal R, Ebinger M (2006) Determining soil quality indicators by factor analysis. *Soil Tillage Res* 87:194–204
- Soil Survey Staff (2014) *Keys to soil taxonomy*, 12th edn. USDA-Natural Resources Conservation Service, Washington
- Souza CMPD, Thomazini A, Schaefer CEGR, Veloso GV, Moreira GM, Fernandes Filho EI (2018) Multivariate analysis and machine learning in properties of ultisols (Argissolos) of Brazilian Amazon. *Rev Bras Ciência do Solo* 42:1–20
- Udo EJ, Ibia TO, Ogunwale JA, Ano AO, Esu IE (2009) *Manual of soil, plant and water analysis*. Sibon Books Publishers Ltd, Nigeria, p 183
- US Geological Survey (2020) *BioData—aquatic bioassessment data for the nation*. U.S. Geological Survey database. Accessed 20 Feb
- Varentsov M, Esau I, Wolf T (2020) High-resolution temperature mapping by geostatistical kriging with external drift from large-eddy simulations. *Mon Weather Rev* 148:1029–1048
- Wallace A (1994) Soil acidification from use of too much fertilizer. *Commun Soil Sci Plant Anal* 25:87–92



- Webster R, Oliver MA (2007) *Geostatistics for environmental scientists*. Wiley, Hoboken
- Wei H, Liu Y, Xiang H, Zhang J, Li S, Yang J (2020) Soil pH responses to simulated acid rain leaching in three agricultural soils. *Sustainability* 12(1):280
- Yan P, Peng H, Yan L, Lin K (2019) Spatial variability of soil physical properties based on GIS and geostatistical methods in the red beds of the Nanxiong Basin, China. *Pol J Environ Stud* 28:2961–2972
- Zeraatpisheh M, Ayoubi S, Sulieman M, Rodrigo-Comino J (2019) Determining the spatial distribution of soil properties using the environmental covariates and multivariate statistical analysis: a case study in semi-arid regions of Iran. *J Arid Land* 11(4):551–566
- Zhang G, Liu F, Song X (2017) Recent progress and future prospect of digital soil mapping: a review. *J Integr Agric* 16(12):2871–2885
- Zhang YY, Wu W, Liu H (2019) Factors affecting variations of soil pH in different horizons in hilly regions. *PLoS ONE* 14:1–13
- Zhu AX, Liu F, Li B, Pei T, Qin C, Liu G, Wang Y, Chen Y, Ma X, Qi F, Zhou CC (2010) Differentiation of soil conditions over low relief areas using feedback dynamic patterns. *Soil Sci Soc Am J* 74(3):861–869





“Gheorghe Asachi” Technical University of Iasi, Romania



## ESTIMATION OF SOIL ORGANIC CARBON DISTRIBUTION BY GEOSTATISTICAL AND DETERMINISTIC INTERPOLATION METHODS: A CASE STUDY OF THE SOUTHEASTERN SOILS OF NIGERIA

John Kingsley<sup>1\*</sup>, Sunday Marcus Afu<sup>2</sup>, Isong Abraham Isong<sup>2</sup>,  
Prince Ageyman Chapman<sup>1</sup>, Ndiye Michael Kebonye<sup>1</sup>, Esther Okon Ayito<sup>2</sup>

<sup>1</sup>Department of Soil Science and Soil Protection, Faculty of Agrobiolgy, Food and Natural Resources,  
Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Praha, Suchdol, Prague, Czech Republic

<sup>2</sup>Department of Soil Science, Faculty of Agriculture, University of Calabar, PMB 1115, Calabar-Nigeria

### Abstract

Soil organic carbon (SOC) plays a significant role in ecosystem protection and sustainable agriculture. The present study aims to estimate the spatial distribution of SOC using three different interpolation methods: ordinary kriging (OK), cokriging (COK), and inverse distance weighting (IDW). Sixty (n = 60) soil samples were collected from the depth of 0–30 cm and analyzed for SOC. The digital elevation model of the site was obtained from USGS explorer at 30 m spatial resolution and processed. Ten (10) terrain attributes were obtained, and a correlation matrix was conducted between SOC and terrain derivatives. The whole dataset was used to evaluate the model accuracy; root mean square error (RMSE) and mean error (ME) were the criteria adopted. Mean value of the SOC of the study area was generally low when compared to the standard rating for tropical soils (< 2%). SOC was significantly (p < 0.01) correlated with LS-factor (r = 0.34\*), negatively correlated with elevation (r = -0.30\*) and profile curvature (r = -0.30\*). IDW performed better (RMSE = 0.75, ME = -0.004) followed by OK (RMSE = 0.78, ME = -0.004) and then COK (RMSE = 0.94, ME = -0.067). Conversely, COK produced the model with the smallest ME with terrain attributes (elevation, LS-factor, and profile curvature). The findings in the study showed that IDW is superior in SOC estimation. COK with the terrain attributes proved to have the capacity as a useful ancillary variable for improving the spatial structure of SOC maps of southeastern Nigeria.

*Key words:* interpolation, kriging, soil organic carbon, tropical soils

*Received:* August, 2020; *Revised final:* November, 2020; *Accepted:* December, 2020; *Published in final edited form:* July, 2021

### 1. Introduction

The importance of estimating spatial soil organic carbon in the biosphere ranges from agricultural productivity to environmental sustainability (Forkuor et al., 2017; Wiesmeier et al., 2014). SOC plays a vital role in sustainable soil fertility, soil quality and wellbeing (Gregorich et al., 1994). SOC controls most soil properties such as porosity, aggregations of particle sizes, moisture retention, and retaining the basic cations in the soil

solution (USDA-NRCS, 1995). The SOC stock of the soils of southeastern Nigeria contributes about 0.2 to 30.8 Mg C ha<sup>-1</sup> to Nigeria's SOC stock (Akpa et al., 2014). The southeastern regions of Nigeria are dominated by agroforestry production, and this agricultural production system can increase carbon stock in the soils through tree biomass under the humid tropical climatic condition. However, there is the challenge of SOC loss which is induced by the adverse effect of climate change (Wiesmeier et al., 2014). SOC content spatially varies over different

\* Author to whom all correspondence should be addressed: e-mail: johnk@af.czu.cz; Phone:+420777871317

agricultural and climatic zones, and there is a need to produce SOC maps for each zone for sustainable agricultural productivity (Liu et al., 2014). More so, quantifying the spatial variability of soil carbon will explain the land ecosystem and establish a baseline for others to calculate the rates of SOC change imposed by management practices (Sanderman and Baldock, 2010).

However, quantifying SOC stocks at a point location is difficult due to the high spatial variability in a given soil unit (Cerri et al., 2000), caused by several soil-forming factors and environmental covariates (Fang et al., 2012). This place demands on the spatial representation of soil organic carbon through regional studies that aids in refining global assessments obtained through regional data (Wang et al., 2010), which is aided through geostatistical and GIS representation (Piccini et al., 2014). This advanced technique emphasizes the benefits of digital soil mapping, which is cost-effective compared to conventional soil mapping in providing soil inventory in formats usable by different soil users. This approach in soil science is referred to as *Pedometrics*, which is a branch of soil science that applies geostatistics, fuzzy membership, pedotransfer functions, and classification trees in soil studies (Mcbratney et al., 2003; Zhu et al., 2010).

Various geostatistical and machine learning techniques have been utilized in the previous to model the spatial distribution of SOC (Kumar et al., 2013). Traditional measures might not make out the spatial allotment of soil properties in the unsampled areas. On the other hand, geostatistics with deterministic models are productive techniques used for examining the spatial differences of soil properties and their irregularity by lessening the fluctuation of evaluation mistake and execution costs (Bhunia et al., 2016). Past studies have utilized geospatial procedures to assess spatial affiliation in soils and to assess soil properties' environmental variability. Besides, more researchers have assessed the expectation exactness of SOC by comparing different modelling approach such as multiple linear regression, random forest, cubist, kriging, inverse distance weighted, empirical Bayesian kriging and so on (Mondal et al., 2016). Mohammad et al. (2010), in their prediction study, stated that ordinary kriging (OK) and cokriging techniques gave better prediction results when compared to the deterministic method [e.g. inverse distance weighting (IDW)] technique for the prediction of the spatial distribution of soil properties. Also, Pang et al. (2011) stated that OK is the foremost common sort of geostatistical technique used in evaluating and modelling surface maps of soil properties.

In spite of the broadly utilized approach in mapping soil properties over the final decades (Zhang et al., 2017), the use of geostatistics techniques and other predictive models to carry soil inventory in Nigeria is constrained (John et al., 2019b). Too, there's small to no evaluated nearby maps in Nigeria. Thus the soaring request for this research for proper soil

management and policymaking. The strategies embraced in this study is due to the reality, there's no particular method that predicts SOC with the leading precision (Mondal et al., 2016).

Southeastern Nigeria's is situated in the humid tropical agro-ecological zone of the country. Soils of the region are highly weathered, dominated by massive sand mixed with low silt and clays fractions (John et al., 2018). Furthermore, in Nigeria, land evaluation and soil nutrient assessment are quite old and outdated. And regardless of the progress in the usage of digital soil mapping (DSM) techniques in regions of the world, little to no research has considered the use of DSM to explain soil nutrient variability in southeastern Nigeria. However, the conventional soil quality assessment method depends on a random soil sampling procedure to acquire an approximated soil fertility status value for a farmer's field (Ayito et al., 2018; Yang et al., 2014). This approach overlooks spatial variability, and the conventional soil laboratory analysis results do not provide randomness of variations obtained from different sampling points. Therefore, some parts of the field may receive excess fertilizer, while others may lack nutrients and experience insufficient productivity levels.

The objective of this study was to estimate SOC distribution using three modelling techniques such as OK and COK and IDW interpolations in soils of southeastern Nigeria.

## 2. Material and methods

### 2.1. Description of the study location

The research was conducted in a consistently steady landscape of Awi in the Akamkpa Area of Cross River State, Nigeria. The research area is situated on 5°22'27.26"N and 8°26'28.39"E for latitude and longitude, respectively (Fig. 1). The site's size is approximately 71.9 hectares on about 180 m high terrain above mean sea level (AMSL). "The area's rainfall and relative humidity ranged between 1500 to 3500 mm and 80 to 90% per year, while the mean annual temperature ranged from 25.4 to 27.5°C (NiMet, 2015)". "The vegetation of the study area is predominantly secondary forest re-growth. Lithologically, the Awi area is underlain by about 40% of the sedimentary basins, occupying roughly 10,000 km<sup>2</sup> of Southeast States (Ekwueme et al., 1990)". According to John et al. (2019), the soils of the area are high in sand, but low in silts and clay contents. "Taxonomically, the soil order of the site is predominantly Ultisols, and the soil classified as Typic paleudults (Aki et al., 2014; John et al., 2019b).

### 2.2. Soil sampling and laboratory analysis

A total of sixty (n = 60) composite samples were collected through stratified random sampling. Samples were collected at a depth of 0 – 30 cm with the aid of a soil auger.

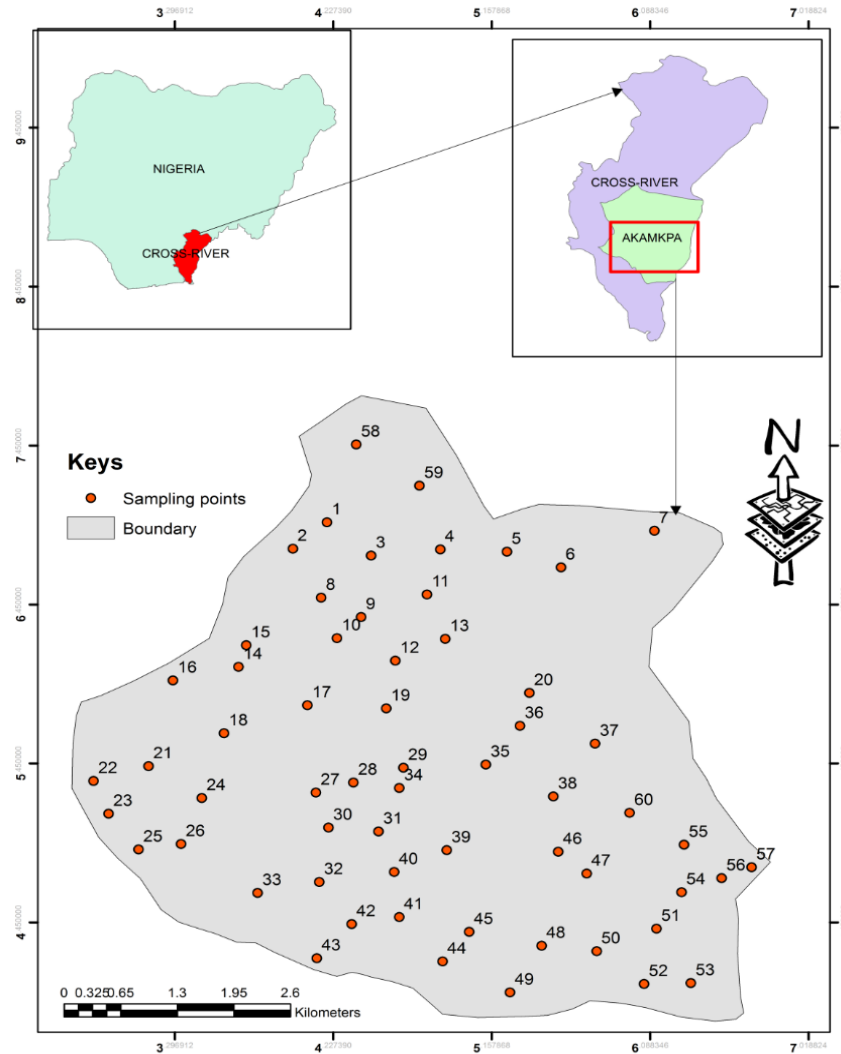


Fig. 1. Map of Awi study site showing the different auger points (n = 60)

Each sample location was labelled and recorded with a hand-held global positioning system (GPS). The samples were taken to the laboratory, air-dried, ground, and sieved with a 0.5 mm mesh. SOC was determined by the standard Walkley-Black wet oxidation method using acid dichromate ( $K_2Cr_2O_7$ ) solution, as outlined in (Udo et al., 2009). This analysis was carried out at the University of Calabar Soil Science Laboratory, as presented in (Eq. 1).

$$\%SOC = N(V1 - V2)0.3f/w \tag{1}$$

where:  $N$  = Normality of  $K_2Cr_2O_7$  solution;  $V1$  = ml ferrous ammonium sulphate required for the blank;  $V2$  = ml ferrous ammonium sulphate needed for the sample;  $w$  = sample in 1 gram.

### 2.3. Terrain model

Digital elevation model (DEM) was obtained from Shuttle Radar Topography Mission (SRTM) at the resolution of 30 x 30 m from and processed in

SAGA-GIS (Olaya, 2004). "The following terrain attributes were obtained, analytical hillshading (Ah), slope (S), aspect (As), plan curvature (Plan C), profile curvature (Profile C), convergence index (CI), topographic wetness index (TWI), LS factor (LS-F), channel network base level, channel network distance (CND), valley depth (VD) and relative slope position (RSP).

### 2.4. Correlation between SOC and terrain attributes

The Pearson correlation coefficient (PCC) is one of the most established effect-size indicators, in part because of its role as a validity coefficient (Morris, 2007). It takes values between the range of -1 to +1, all-encompassing, and yields a measure of the strength of the linear relationship that exists between two variables. Furthermore, for the purpose of this current study, we only considered terrain attributes that showed a significant correlation ( $p < 0.001, 0.01, 0.1$ ) with SOC and observed to influence its variability in the study location. These terrain attributes were incorporated into the COK model.



## 2.5. Spatial modelling for estimating soil organic carbon

### 2.5.1. Geostatistical technique

The geostatistical method uses unbiased predictions with minimum variance for the targeted soil property (Stein and Corsten, 1991). OK, and COK is among the various types of geostatistical methods. The OK process uses an estimated mean of a particular soil property at a known location to predict the value at an unsampled location (Bishop and McBratney, 2001; Goovaerts, 1997; Grunwald et al., 2008) (Eq. 2), whereas COK uses information on several variable types to predict a particular target variable (in this case SOC). And these variables must exhibit a strong relationship with the targeted property (Bivand et al., 2008; Tziachris et al., 2017).

$$Z'(x_0) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (2)$$

where:  $Z'(x_0)$  is the predicted/interpolated value for point  $x_0$ ,  $Z(x_i)$  is the known value, and  $\lambda_i$  is the kriging weight for the  $Z(x_i)$  values. It can be calculated by the semi-variance function of the variables on the condition that the estimated value is unbiased and optimal (Eq. 3).

$$\gamma(h) = 1/2N(h) \sum_{i=1}^n [Z(x_i) - Z(x_i + h)]^2 \quad (3)$$

where:  $\gamma(h)$  is the semi-variance,  $N(h)$  is the point group number at distance  $h$ ,  $Z(x_i)$  is the numerical value at position  $x_i$ , and  $Z(x_i + h)$  is the numerical value at a distance  $(x_i + h)$ .

### 2.5.2. Deterministic technique

IDW is a deterministic predictive tool that determines cell values using a linearly weighted combination of a set of sample points and where the weight is a function of inverse distance (Philip and Watson, 1982; Bhunia et al., 2016; Liu et al., 2017). Estimated values were interpolated based on the data from surrounding locations using the Eqs. (4-5).

$$Z(x_0) = \sum_{i=1}^n w_i Z(x_i) \quad (4)$$

where:  $Z(x_0)$  is the estimated value,  $w_i$  is the weight assigned to the value at each location  $Z(x_i)$ ,  $n$  is the number of close neighbouring sampled data points used for estimation.

The weights were estimated using Eq. (5):

$$w_i = 1/d_i^p / \sum_{i=1}^n 1/d_i^p \quad (5)$$

where:  $d_i$  is the distance between the estimated point and the sample point,  $p$  is an exponent parameter.

## 2.6. Model validation of the spatial soil organic carbon estimation

In the evaluation of our spatial estimation, we used the total data to estimate the trend and

autocorrelation of our models. "The interpolated result was then extracted to the whole data points. Root mean square error (RMSE) and mean error (ME). The RMSE gives an estimate of the standard deviation of the residuals (prediction errors)." While mean error (ME) is taken as the mean of residuals, it calculates the deviation of the predicted value Eqs. (6-7) expresses them as:

$$RMSE = \sqrt{1/n \sum_{i=1}^n (p_i - o_i)^2} \quad (6)$$

$$ME = 1/n \sum_{i=1}^n (p_i - o_i) \quad (7)$$

where:  $p_i$  = predicted values,  $o_i$  = observed values,  $n$  = the number of validation points. Interpretatively, a good model should have a low RMSE and ME close to 0 if the predicted results are unbiased (Robinson and Metternicht, 2006).

## 2.7. Data analysis

SOC spatial maps were produced via ArcGIS. Terrain attributes were derived through System for Automated Geoscientific Geographical Information System (SAGA-GIS) software. At the same time, discrete statistics and estimate the correlation matrix between SOC and terrain attributes processed via R studio.

## 3. Results and discussion

### 3.1. Descriptive statistics

The samples summary statistics of SOC and terrain attributes are presented in Tables 1-2, respectively. The result revealed that the SOC value of the area ranged from 0.7–3.2%, with a mean of 1.77%. SOC was very low when compared with Landon (1991) rating for tropical soils. The result obtained here is similar to the report of John and Akpan-Idiok (2019b) and in contrast with that Abua and Eyo (2013) and Aki et al. (2014). They rated moderate SOC in similar soils. Furthermore, the low SOC obtained in this study may be attributed to surface runoff (Larsen et al., 2014), high temperature and precipitation (Bolliger et al., 2006), increased soil acidity (John et al., 2019a) and intensive cropping without adequate nutrient management (Ayito et al., 2018). The measured SOC expressed a normal distribution with high variability (CV=37.8), a positive skewness of 0.39, and a kurtosis of 2.15. On the other hand, the terrain attributes showed a normal distribution and were not transformed as well. However, LS-F, Profile C, CND and VD produced high variability with CV values of 37.8, 38, 928.9, 37.62 and 37.56, respectively, compared to the standard rating outlined by Gubiani et al. (2011). Simultaneously, RSP and Elev yielded moderate and low variability with CV values of 22.85 and 3.3, respectively.

**Table 1.** Descriptive statistics of SOC

Variables	Mean	Min	Max	SD	CV	Skewness	Kurtosis	Data Transformation
SOC (%)	1.77	0.7	3.2	0.67	37.8	0.39	2.15	None

**Table 2.** Descriptive statistics of some selected terrain attributes

	Elev (m)	LS-F	RSP	Profile C	CND	VD
Mean	165.9	2.10	0.50	651383.04	13.37	13.26
Standard Deviation	5.46	0.48	0.19	6050954.93	5.03	4.98
Kurtosis	2.30	4.55	-0.51	-0.49	-0.45	-0.53
Skewness	0.19	1.99	0.32	0.02	0.35	-0.31
Minimum	155.20	1.55	0.10	-14546609.78	2.68	1.54
Maximum	178.41	3.91	0.94	14363443.95	25.40	23.87
CV	0.19	22.85	38	928.90	37.62	37.56
Confidence Level(95%)	1.41	0.12	0.05	1563127.41	1.30	1.29
Data Transformation	None	None	None	None	None	None

Elev = Elevation; LS-F = LS-factor; RSP = Relative Slope Position Profile C= Profile Curvature; CND = Channel Network Distance; VD = Valley Depth

Generally, the variables were employed untransformed for the modelling purpose.

3.2. Correlation between SOC and terrain attributes

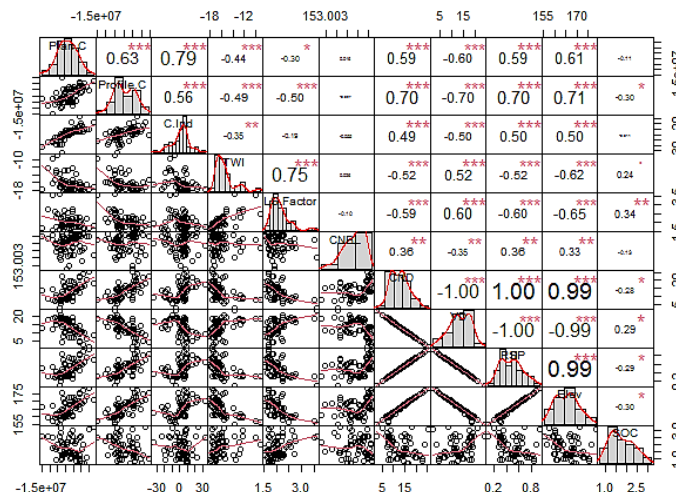
A Pearson correlation analysis was estimated to explain the relationship between SOC with the terrain attributes (Fig. 2). The result revealed that SOC was negative and significantly ( $p < 0.01$ ) correlated with Elev ( $r = -0.30^*$ ), RSP ( $r = -0.29^*$ ), CND ( $r = -0.29^*$ ), Profile C ( $r = -0.30^*$ ) but positively and significantly correlated with LS-factor ( $r = 0.34^*$ ). The result further revealed that LS-factor was the highest terrain attributes that yielded the highest correlation with SOC compared to other terrain attributes. The negative and significant ( $p < 0.01$ ) correlation obtained between SOC and Elev is similar to the report by Kozłowski and Komisarek (2018).

Also, the same report was not consistent with the result obtained for SOC and Profile C in our study. Furthermore, the result of our study corroborates with that of Li et al. (2018), who observed significant correlations between SOC and LS-factor, Profile C, and other terrain derivatives. In this study, the Pearson correlation coefficient presented the relationship

between SOC and terrain attributes. It revealed the capability of estimating SOC variability via terrain attributes. In the COK modelling, terrain attributes with relatively high correlation were used. These terrain attributes include LS-F, Elev, and Profile C as they could improve the prediction of SOC OC in the local landscape of southeastern Nigeria.

3.3. Spatial estimation of SOC

In this present study, OK, COK and IDW methods were used to estimate the spatial variability of SOC. Discrete statistics of the interpolation output is presented in Table 2, while the fitted semivariograms for the OK and COK model are shown (Fig. 3a -b). The Semivariogram model revealed that OK and COK produced a stable model. OK was fitted with nugget = 0.19, sill = 0.42 and range = 1.998 while COK was fitted with nugget = 0.28, sill = 0.30 and range = 1.997. On the other hand, COK presented a high nugget effect (0.28) compared to OK (0.19). Elev, LS-F, and Profile C may have contributed to this variation as they have been reported to influence SOC spatial variability (Wu et al., 2009; Tsui et al., 2013).



**Fig. 2.** Correlation matrix of SOC and terrain derivatives

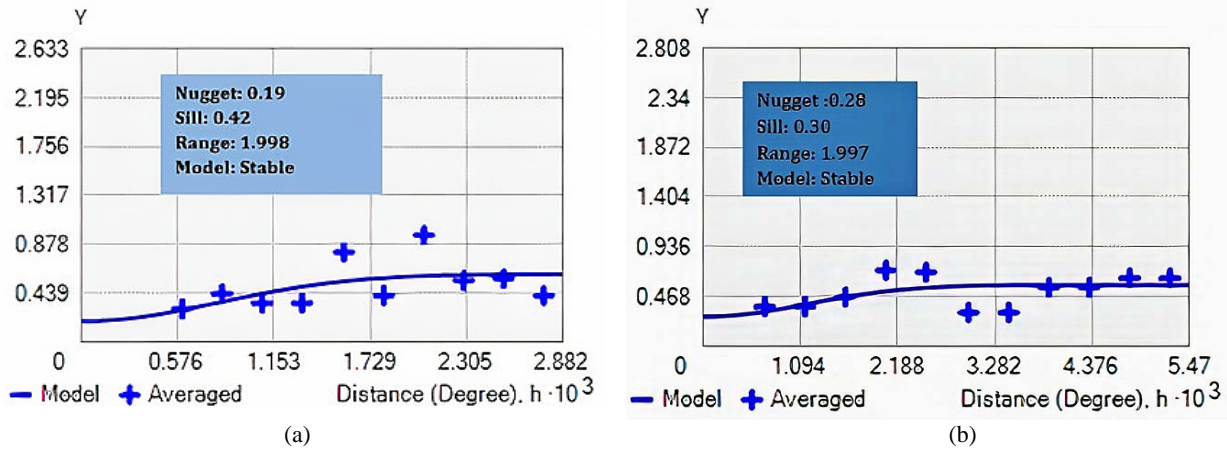


Fig. 3. (a) OK semivariogram (b) COK semivariogram

The spatial autocorrelation for OK and COK was 31.1% and 48.2%, respectively. Spatial autocorrelation is the nugget to sill ratio as defined by Cambardella et al. (1994). The values obtained for OK and COK showed that the models gave a moderate spatial autocorrelation as they fell within (> 25% < 75%), a criterion by Cambardella et al. (1994). The variation of SOC seen in the site may be associated with the accumulation of mineral and organic material from relative slope positions, as suggested by Brodsky et al. (2013).

3.4. Comparison of OK, COK and IDW interpolation

In evaluating the model with the best performance, the whole dataset was employed. The criteria for the best model was a low RMSE and ME value (Yang et al., 2009). As shown in Table 3, OK (RMSE = 0.78, ME = -0.004), COK (RMSE = 0.94, ME = -0.067) and IDW (RMSE = 0.75, ME = -0.004). The results revealed that the ME values of the three methods were close to 0, indicating that predicted values were unbiased. Furthermore, the cross-validation result presented in Table 2 revealed that IDW was more accurate than both OK and COK having the lowest RMSE value. The OK model followed the IDW as the next model with a low RMSE. IDW as the best model agrees with Li and Heap (2008) and contrasts with Bhunia et al. (2016). COK, on the other hand, yielded a smaller mean error compared to OK and IDW. The narrow mean error obtained may be attributed to the added terrain attributes (Elev, LS-factor and profile C).

COK model also suggests that terrain attributes could serve as excellent auxiliary variables for improving the reliability of spatial SOC prediction.

The result obtained here is similar to the report by Yang et al. (2014), who reported the importance of elevation and slope in estimating SOC variability in Southwest China. Also, Triantafyllis et al. (2001), Wu et al. (2009), Tziachris et al. (2017), and Saleh (2018) reported a low mean error for COK. Besides that, Chabala et al. (2017) and Bhunia et al. (2016) reported OK as the best model for SOC prediction in their studies. Nevertheless, comparative interpolation studies of SOC prediction have always shown mixed results, often associated with available data and the type of interpolation technique (Chabala et al., 2017).

3.5. Prediction maps of SOC by the different interpolation methods

SOC predicted maps using OK, COK and IDW models are presented in Fig. 4. The maps structures showed significant differences, revealing a high spatial variability in SOC. The map developed from OK was smoother than that produced from COK and IDW, respectively.

COK, as well as IDW, revealed more details in local areas as compared. The result obtained in the predicted map of OK corroborates with the report by Wu et al., (2009), who reported a smooth trend in the OK map of soil organic matter.

The predicted SOC map by OK was less spatially detailed (i.e. evenly distributed) than that by COK and IDW in some local regions, such as the central part in the study site, as shown in the SOC prediction maps (Fig. 4 (a-c)). SOC ranged from 0.98-2.64%, 1.18-2.32% and 0.70-3.2% in OK, COK and IDW maps, respectively. Generally, the predicted SOC maps revealed that SOC was relatively high in the central part of the research area.

Table 3. Comparison of the interpolation methods to map SOC distribution

Interpolation methods	RMSE (%)	ME
OK	0.78	-0.004
COK	0.94	-0.067
IDW	0.75	-0.004

OK: Ordinary kriging; COK: Cokriging; IDW: Inverse distance weighting

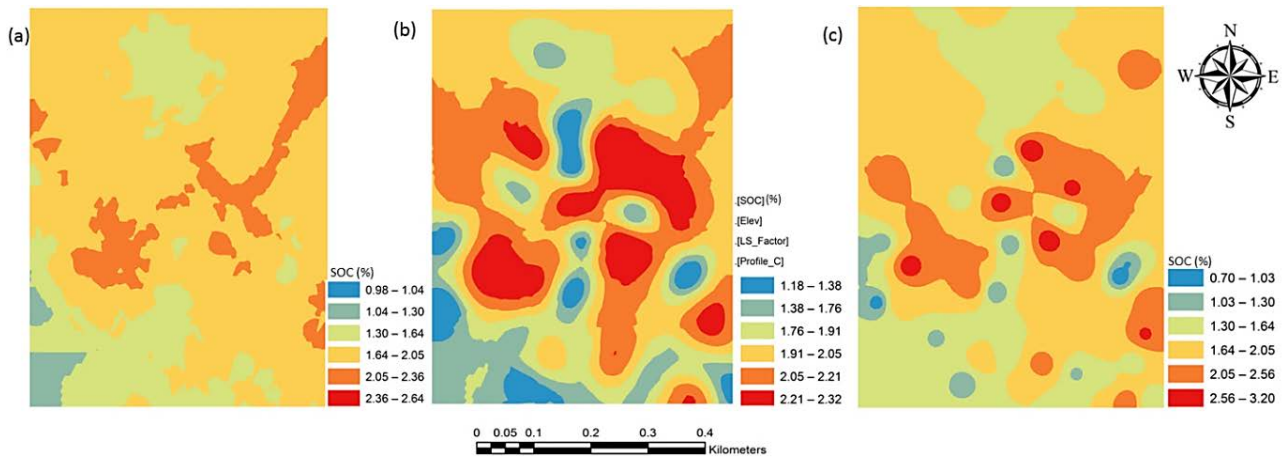


Fig. 4. SOC(%) prediction maps via (a) OK model (b) COK model (c) IDW model

### 3.6. Descriptive statistics of predicted soil organic carbon (SOC)

The summary statistics of the predicted SOC by the three different models are presented in Table 4. The predicted SOC values also presented a normal distribution for the interpolation methods. SOC predicted value was 1.68% for OK, COK, and IDW, respectively. Also, the measure SOC minimum and maximum value were the same as that of IDW prediction.

The descriptive statistics of the predicted SOC values showed a normal distribution like the measured SOC. The result is supported by the report of Chabala et al. (2017). Despite that, the work revealed that SOC predicted was lower than SOC measured value. And when compared to Landon (1991) ratings, predicted SOC was observed to be very low (< 2 %). This shows that this level of SOC cannot sustain an intensive cropping system in the area. The result obtained here may be attributed to lumbering activities often carried out in the area.

Table 4. Predicted SOC using OK, COK and IDW

	Mean	Min	Max	SD	CV	skewness	kurtosis
OK	1.68	0.98	2.64	0.27	16.1	0.54	-0.05
COK	1.68	1.18	2.32	0.37	22	0.33	-0.33
IDW	1.68	0.7	3.2	0.52	31	0.88	0.91

This action results in significant losses of SOC, which tend to reduce further crop yields under continuous cultivation. This act of deforestation would further lead to the decline in soil fertility through increased soil erosion, reduction of litter influx after canopy removal and boosted decomposition and nutrient mineralization rates after forest clearance.

## 4. Conclusions

In this present study, OK, COK and IDW interpolations were performed and compared to evaluate the accuracy of our prediction of the geographical variability SOC.

The study revealed that SOC was generally low in the research site. SOC demonstrated a moderate spatial dependence and explained the essence of estimating SOC spatial variability in southeastern Nigeria. Among the three interpolations, IDW was the best performing model. At the same time, the COK model gave the smallest mean error, which was observed to have occurred due to terrain attributes. The predicted SOC map by COK with Elev, LS-F and profile C covariates improved the OK and IDW maps, respectively. The COK map was more detailed, showing the capability of terrain attributes being robust ancillary variables for improving detailed spatial SOC maps.

In conclusion, the SOC created maps by COK and IDW of the study area could be adopted by both soil and land users to help grow different crops concerning their different nutrient needs for adequate agricultural production management. Besides that, the created maps could be used as a reference point for various soil purposes, ranging from sampling optimization to updating soil maps with more ancillary variables. Furthermore, for future studies, it is recommended that different auxiliary covariates be introduced and an increase in sample density to improve the accuracy of the models in estimating SOC.

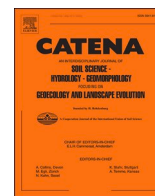
## Acknowledgements

Mr Kingsley John Funding This study was supported by an internal PhD grant no. SV20-5-21130 of the Faculty of Agrobiological, Food and Natural Resources of the Czech University of Life Sciences Prague (CZU). Secondly, the Czech Science Foundation projects no. 17-277265 (Spatial prediction of soil properties and classes based on position in the landscape and other environmental covariates) and 18-28126Y (Soil contamination assessment using hyperspectral orbital data) for the financial support. Thirdly, the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre) and the European project no. CZ.02.1.01/0.0/0.0/16\_019/0000845.

## References

- Abua M., Eyo E., (2013), Assessment of soils around quarry terrain in Akamkpa local government area, Cross River State-Nigeria, *Merit Research Journal of Agricultural Science and Soil Sciences*, **1**, 001-005.
- Aki E.E., Esu I.E., Akpan-Idiok A.U., (2014), Pedological study of soils developed on Biotite- Hornblende-Gneiss in Akamkpa Local Government Area in Cross River State, Nigeria, *International Journal of Agricultural Research*, **9**, 187-199.
- Akpa S.I.C., Odeh I.O.A., Bishop T.F.A., Hartemink AE, (2014), Digital mapping of soil particle-size fractions for Nigeria, *Soil Science Society of America Journal*, **78**, 1953-1966.
- Ayito E.O., Iren O.B., John K., (2018), Effects of neem-based organic fertilizer, NPK and their combinations on soil properties and growth of Okra (*Abelmoschus esculentus*) in a Degraded Ultisol of Calabar, Nigeria, *International Journal of Plant & Soil Science*, **25**, 1-10.
- Brodsky L., Vasat R., Klement A., Zadorova T., Jaksik O., (2013), Uncertainty propagation in VNIR reflectance spectroscopy of soil organic carbon mapping, *Geoderma*, **199**, 54-63.
- Bhunja S.G., Shit PK, Maiti R., (2016), Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC), *Journal of the Saudi Society of Agricultural Sciences*, **17**, 114-126.
- Bivand R.S., Pebesma E.J., Gomez-Rubio V., (2008), *Applied Spatial Data Analysis with R*, Springer, New York.
- Bishop T.F.A., McBratney A.B., (2001), A comparison of prediction methods for the creation of field-extent soil property maps, *Geoderma*, **103**, 149-160.
- Bolliger A., Magid J., Amado J.C.T., Neto F.S., dos Santos Ribeiro M.D.F., Calegari A., Ralisch R., de Neergaard A., (2006), Taking stock of the Brazilian "Zero - Till Revolution": A review of landmark research and farmers' practice, *Advances in Agronomy*, **91**, 47-110.
- Cambardella C., Moorman T., Parkin T., Karlen D., Novak J., Turco R., Konopka A., (1994), Field-scale variability of soil properties in central Iowa soils, *Soil Science Society of America Journal*, **58**, 1501-1511.
- Cerri C.C., Bernoux M., Arrouays D., Feigl B.J., Piccolo M.C., (2000), *Carbon Stocks in Soils of the Brazilian Amazon*, In: *Global Climate Change and Tropical Ecosystems*, Lal R., Kimble J.M., Stewart B.A., (Eds), CRC Press, Boca Raton, 33-50.
- Chabala L.M., Mulolwa A., Lungu O., (2017), Application of ordinary kriging in mapping soil organic carbon in Zambia, *Pedosphere*, **27**, 338-343.
- Fang X., Xue Z., Li B., An S., (2012), Soil organic carbon distribution in relation to land use and its storage in a small watershed of the Loess Plateau, China, *Catena*, **88**, 6-13.
- Forkuor G., Hounkpatin O.K.L., Welp G., Thiel M., (2017), High-resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models, *PLoS One*, **12**, 0170478, <https://doi.org/10.1371/journal.pone.0170478>.
- Goovaerts P., (1997), *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, USA.
- Gregorich E.G., Carter M.R., Angers D.A., Monreal C.M., Ellert B.H., (1994), Toward minimum data set to assess soil organic matter quality in agricultural soils, *Canadian Journal of Soil Science*, **74**, 885-901.
- Grunwald S., Osborne T.Z., Reddy K.R., (2008), Temporal trajectories of phosphorus and pedo-patterns mapped in Water Conservation Area 2, Everglades, Florida, USA, *Geoderma*, **146**, 1-13.
- Gubiani P.I., Reichert J.M., Reinert D.J., Gelain N.S., (2011), Implications of the variability in soil penetration resistance for statistical analysis, *Revista Brasileira de Ciência do Solo*, **35**, 1491-1498.
- John K., Ayito E.O., Odey S., (2018), Interaction between some soil physicochemical properties and weather variables on sub-humid tropical rainforest soils of cross river state, Southeastern Nigeria, *Annual Research & Review in Biology*, **29**, 1-12.
- John K., Solomon O.L., Ayito E.O., Ndiye M.K., Ogeh J.S., Vít P., (2019a), Predictive mapping of soil properties for precision agriculture using geographic information system (GIS) based geostatistics models, *Modern Applied Science*, **13**, 60-74.
- John K., Akpan-Idiok A.U., (2019b), Land evaluation, characterization and classification of soil for the proposed oil palm plantation in Ekpri Ibami, Akamkpa Local Government Area, Nigeria, *International Journal of Environment Agriculture and Biotechnology*, **4**, 621-634.
- Kumar S., Lal R., Liu D., Rafiq R., (2013), Estimating the spatial distribution of organic carbon density for the soils of Ohio, USA, *Journal of Geographical Sciences*, **23**, 280-296.
- Kozłowski M., Komisarek J., (2018), Influence of terrain attributes on organic carbon stocks distribution in soil toposquences of central Poland, *Soil Science Annual*, **69**, 215-222.
- Landon J.R., (1991), *Booker Tropical Soil Manual*, Longman Scientific and Technical Essex, UK.
- Larsen E., Grossman J., Edgell J., Hoyt G., Osmond D., Hu S., (2014), Soil biological properties, soil losses and corn yield in long-term organic and conventional farming systems, *Soil Tillage Research*, **139**, 37-45.
- Li X., McCarty G., Karlen D., Cambardella C., (2018), Topographic metric predictions of soil redistribution and organic carbon in Iowa cropland fields, *Catena*, **160**, 222-232.
- Li J., Heap A.D., (2008), A review of spatial interpolation methods for environmental scientists, Canberra, Australia: Geoscience Australia, Record 2008/23, 137-154.
- Liu L., Wang H., Dai W., Lei X., Yang X., Li X., (2014), Spatial variability of soil organic carbon in the forestlands of northeast China, *Journal of Forest Research*, **25**, 867-876.
- Liu W., Hai-Rong Z., Da-Peng Y., Sheng-Li W., (2017), Adaptive Surface Modeling of Soil Properties in Complex Landforms, *ISPRS International Journal of Geo-Information*, **6**, 178, <https://doi.org/10.3390/ijgi6060178>.
- McBratney A.B., Mendoça-Santos M.L., Minasny B., (2003), On digital soil mapping, *Geoderma*, **117**, 3-52.
- Mohammad, Z.M., Taghizadeh-Mehrjardi R., Akbarzadeh, A., (2010), evaluation of geostatistical techniques for mapping spatial distribution of soil pH, salinity and plant cover affected by environmental factors in Southern Iran, *Notulae Scientia Biologicae*, **2**, 92-103.
- Mondal A., Khare D., Kundu, S., Mondal S., Mukherjee S., Mukhopadhyay A., (2016), Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data, *The Egyptian Journal of Remote Sensing and Space Sciences*, **20**, 61-70.
- Morris S.B., (2007), *Book Review: Hunter J.E., Schmidt F.L., (2004), Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, 2nd Edition,*

- Thousand Oaks, CA: Sage, *Organizational Research Methods*, **11**, 184-187.
- NiMet, (2015), Annual meteorological reading at Calabar Station, Nigerian Meteorological Station, On line at: <https://www.nimet.gov.ng/>.
- Olaya V., (2004), A Gentle Introduction to SAGA GIS, On line at: <https://freecomputerbooks.com/A-Gentle-Introduction-to-SAGA-GIS.html>.
- Pang S., Li T.X., Zhang X.F., Wang Y.D., Yu H.Y., (2011), Spatial variability of cropland lead and its influencing factors: A case study in Shuangliu County, Sichuan province, China, *Geoderma*, **162**, 223-230.
- Philip G.M., Watson D.F., (1982), A precise method for determining contoured surfaces, *Australian Petroleum Exploration Association Journal*, **22**, 205-212.
- Piccini C., Marchetti A., Francaviglia R., (2014), estimation of soil organic matter by geostatistical methods: use of auxiliary information in agricultural and environmental assessment, *Ecological Indicator*, **36**, 301-314.
- Robinson T.P., Metternicht G., (2006), Testing the performance of spatial interpolation techniques for mapping soil properties, *Computer Electronics and Agriculture*, **50**, 97-108.
- Saleh A.M., (2018), Spatial variability mapping of some soil properties in Jadwal Al\_Amir Project/Babylon/Iraq, *Journal of the Indian Society of Remote Sensing*, **46**, 1481-1495.
- Sanderman J., Farquharson R., Baldock J.A., (2010), *Soil Carbon Sequestration Potential*, A review for Australian agriculture, CSIRO Sustainable Agriculture Flagship Report, prepared for Department of Climate Change and Energy Efficiency, On line at: [www.csiro.au/resources/Soil-Carbon-Sequestration-Potential-Report.html](http://www.csiro.au/resources/Soil-Carbon-Sequestration-Potential-Report.html).
- Stein A., Corsten LCA, (1991), Universal kriging and cokriging as regression procedure, *Biometrics*, **47**, 575-587.
- Triantafyllis J., Odeh I.O.A., Mcbratney A.B., (2001), Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton, *Soil Science Society of America Journal*, **65**, 869-878.
- Tsui C., Chen-Chi T., Zueng-Sang C., (2013), Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan, *Geoderma*, **210**, 119-127.
- Tziachris P., Eirini M., Frantzis P., Maria P., (2017), Spatial modelling and prediction assessment of soil iron using kriging interpolation with pH as auxiliary information, *ISPRS International Journal of Geo-Information*, **6**, 283, <https://doi.org/10.3390/ijgi6090283>.
- Udo E.J., Ibia T.O., Ogunwale J.A., Ano AO, Esu IE, (2009), *Manual of Soil, Plant and Water Analysis*, Sibon Books Publishers Ltd, Nigeria.
- USDA-NRCS, (1995), Soil Survey Laboratory Information Manual, Soil Survey Investigations Report No. 45. National Soil Survey Center, Soil Survey Laboratory, United States Department of Agriculture-Natural Resources Conservation Service, Lincoln, Nebraska.
- Wiesmeier M., Barthold F., Spörlein P., Geuß U., Hangen E., Reischl A., (2014), estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany), *Geoderma Regional*, **1**, 67-78.
- Wang Y., Fu B., Yi H.L., Song C., Luan Y., (2010), Local-scale spatial variability of soil organic carbon and its stock in the hilly area of the Loess Plateau, *China Quaternary Research*, **73**, 70-76.
- Wei J.B., Xiao D.N., Zeng H., Fu Y.K., (2008), Spatial variability of soil properties in relation to land use and topography in a typical small watershed of the black soil region, northeastern China, *Environmental Geology*, **53**, 1663-1672.
- Wu C., Jiaping W., Yongming L., Limin Z., Stephen D., (2009), Spatial prediction of soil organic matter content using cokriging with remotely sensed data, *Soil Science Society of America Journal*, **73**, 1202-1209.
- Yang Q., Jiang Z., Ma Z., Li H., (2014), Spatial prediction of soil water content in karst area using prime terrain variables as auxiliary cokriging variable, *Environmental Earth Sciences*, **72**, 4303-4310.
- Yang Y., Zhu J., Tong X., Wang D., (2009), *Spatial Pattern Characteristics of Soil Nutrients at the Field Scale*, Li D., Zhao C. (Eds.), *Computer and Computing Technologies in Agriculture*, Springer, Berlin, Heidelberg, 125-134.
- Zhang G., Liu F., Song X., (2017), Recent progress and future prospect of digital soil mapping: A review, *Journal of Integrative Agriculture*, **16**, 2871-2885.
- Zhu A.X., Liu F., Li B.L., Pei T., Qin C.Z., Liu G.H., Wang Y.J., Chen Y.N., Ma XW, Qi F., Zhou C.H., (2010), Differentiation of soil conditions over flat areas using land surface feedback dynamic patterns extracted from MODIS, *Soil Science Society of America Journal*, **74**, 861-869.



## Hybridization of cokriging and gaussian process regression modelling techniques in mapping soil sulphur

Kingsley John<sup>a,\*</sup>, Prince Chapman Agyeman<sup>a</sup>, Ndiye Michael Kebonye<sup>a</sup>, Isong Abraham Isong<sup>b</sup>, Esther O. Ayito<sup>b</sup>, Kokei Ikpi Ofem<sup>b</sup>, Cheng-Zhi Qin<sup>c</sup>

<sup>a</sup> Department of Soil Science and Soil Protection, Faculty of Agrobiological, Food, and Natural Resources, Czech University of Life Sciences, Kamýcká 129, 16500 Prague, Czech Republic

<sup>b</sup> Department of Soil Science, University of Calabar, Nigeria

<sup>c</sup> State Key Laboratory of Resources & Environmental Information System, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, PR China

### ARTICLE INFO

#### Keywords:

Soil nutrient distribution  
Digital soil mapping  
Machine learning  
Cokriging  
Agricultural productivity

### ABSTRACT

As a widely used soil mapping method, the kriging method involves a high sampling point to generate quality and accurate maps. Combining kriging and machine learning (ML) can produce soil maps with fewer number sampling points. This study's objective was to implement a hybrid approach based on the Cokriging (Cok) and an ML technique [i.e., Gaussian process regression (GPR)]. The hybrid method (called the Cok-GPR method) uses the Cok (Coki,  $i = 1$  to  $n$ ) as a predictor method of the soil sulphur and then uses GPR to improve the prediction accuracy. The proposed method was compared with the Cok and the GPR models, respectively, in a case study. Soil samples ( $n = 115$ ) were collected from the topsoil (0–20) at the agricultural site of approximately 889.8 km<sup>2</sup> size. S, Ca, K, Mg, Na, P, and V were estimated via Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) equipment and presented as S\_ICP-OES (response variable), and predictors (Ca\_ICP-OES, K\_ICP-OES, Mg\_ICP-OES, Na\_ICP-OES, P\_ICP-OES, and V\_ICP-OES), respectively. For GPR and Cok-GPR, an 80% (calibration) to 20% (validation) random dataset split was performed. The calibration dataset was implemented under  $k = 10$ -fold cross-validation, repeated five times. All the models were evaluated by MAE, RMSE and R2 criteria. According to the model and map performances. Cok1 model via Ca\_ICP-OES, K\_ICP-OES, Mg\_ICP-OES gave the best model (MAE = -1.28 mg/kg, RMSE = 164.42 mg/kg, R2 = 0.85). Its corresponding GPR1 approach, modelled with the same predictors produced the best (MAE = 85.43 mg/kg, RMSE = 137.59 mg/kg, R2 = 0.83). While the hybrid Cok1-GPR model produced MAE = 76.84 mg/kg, RMSE = 102.11 mg/kg, and R2 = 0.91. The model outperformed both the Cok and GPR models, respectively. The proposed Cok-GPR model can be applied to efficiently predict soil nutrient element levels at the regional level and be useful during policymaking.

### 1. Introduction

Sulphur (S) is essential for the growth and development of crops. For more than a century, the essentiality of S has been recognized (Sager, 2012), and recently, it has received more attention compared with the primary soil nutrients, N, P, and K (Sager, 2012). This is due to its role in good crop production and plant nutrition. Plants require S to synthesize essential amino acids and proteins (e.g. methionine and cysteine), vitamins and coenzymes, glucoside oils, structurally and physiologically important disulfide linkages and sulphhydryl groups, as well as in the activation of certain enzymes (Lucheta and Lambais, 2012). Sulphur is

present in all soils and could be derived from parent rock materials, atmospheric deposition, marine aerosols, industrial gases, gases/particulates, volcanic eruptions and fertilizer formulations (Sager, 2012). Stevenson (1986) reported that total S content of soils varies over a wide range, from as little as 20 mg kg<sup>-1</sup> in highly weathered soils in humid regions to over 50,000 mg/kg in calcareous and saline soils of arid semiarid areas. For example, Olson and Englestad (1972) provided an excellent summary of average topsoil values for total S in the temperate zone: 500 mg/kg for Mollisols, 400 mg/kg for Alfisols and 200 mg/kg for Ultisols. Similarly, a Eutrudand from Hawaii contained 1280 mg/kg S in its topsoil (Fox et al. 1971). In Brazil, a clayey Oxisol under native

\* Corresponding author.

E-mail address: [johnk@af.czu.cz](mailto:johnk@af.czu.cz) (K. John).

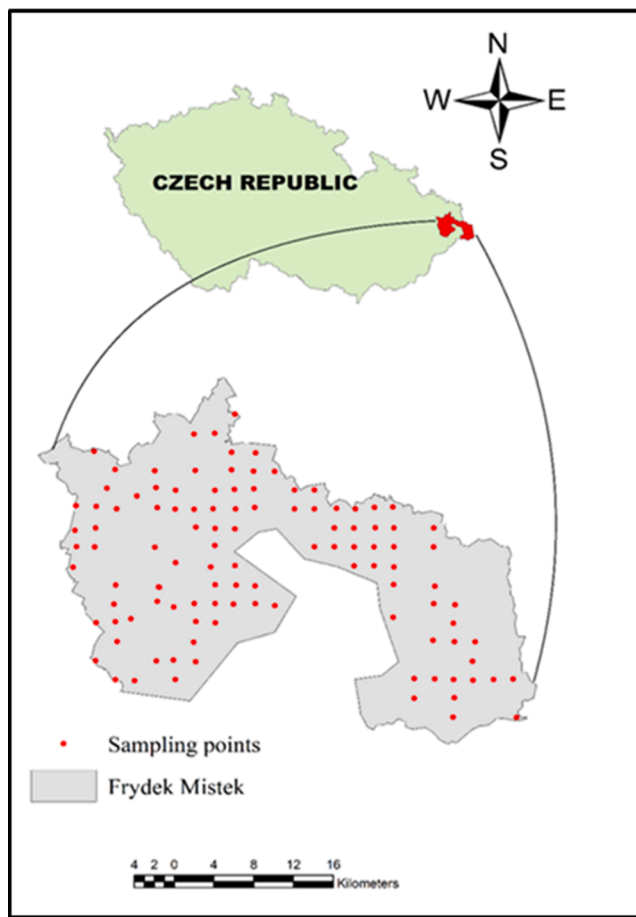


Fig. 1. Map of the study area showing sampling points.

savannah had 251 mg/kg S and a sandy Ultisol only 40 mg/kg S (McClung et al. 1959).

In precision agriculture (PA), digital soil mapping (DSM) has become a crucial step towards accurate and proper soil and crop management procedure and policymaking. Digital soil mapping involves applying mathematical and statistical techniques to estimate a targeted soil property via environmental variables (e.g. covariates) known to influence the targeted properties spatial distribution (McBratney et al., 2003). In the past, the approach involved applying geostatistics and other interpolations, interpolating unknown values from neighbouring points sampled (Hengl et al., 2004; Hedge et al., 2017). There are several interpolation methods to predict values at unsampled positions. Inverse Distance Weighting (IDW), Ordinary Kriging (OK), Universal kriging (UK), Cokriging (Cok) as well as others, have been widely applied in PA as spatial interpolation techniques used (Zhu et al., 1997; Qin et al., 2021; Hengl et al., 2004; Valente et al., 2012; Coelho et al., 2018; Agyeman et al., 2020; Pei et al., 2010). OK has been reported to be excessive data-dependent, requiring many regularly spaced data points, assuming significant spatial autocorrelation trends (Scull et al., 2003; Sekulić et al., 2020). And it is also applicable to Cok, a stochastic interpolator (Setiyoko et al., 2019). It requires high sampling density and user knowledge to model the semivariogram (Giacomin et al., 2014; Webster and Oliver, 1992). Webster and Oliver, 2001; Pouladi et al., 2019 reported that sample points between 30 and 140 have an excellent semivariance estimation for each specific distance. And collecting larger samples where relative larger areas are concerned may be costly, tedious, and time-consuming, while fewer samples may yield high uncertainty in the prediction. Considering the application of only geostatistics to assess plant needs for S and estimate the S-supplying power of soils, it is imperatives to consider its spatial distribution over a large

geographical extent (e.g. regional scale) concerning the representative samples. Therefore, this would not provide sufficient detailed information about the S variability levels required for sustainable cropping. Hence, S's spatial distribution can be studied using the combination of kriging and machine learning (ML) as done elsewhere (e.g., regression kriging).

Among the several challenges of creating quality and accurate maps, many sampling points are the most prevalent ones necessary to produce accurate maps of targeted soil property. And one approach to solving this problem is proposing techniques that require fewer sampling points to create accurate maps. These innovative proposed approaches should be flexible to use covariates with available high-density observations, such as digital elevation model (DEM) and remote sensing data.

Machine learning (ML) algorithm has recently proved to be an efficient technique in predicting and mapping soil property (Khaledian and Miller, 2020; Kebonye et al., 2020; Kebonye et al., 2021; John et al., 2020). ML is an automated process of learning by algorithms based on data size. ML can accommodate non-linearity and multicollinearity, and they can overcome overfitting with limited soil sample points and environmental covariates and can recognize data patterns (Drake et al., 2006; Gautam et al., 2011; Heung et al., 2016; Liakos et al., 2018; Parmley et al., 2019; Khaledian and Miller, 2020). Several ML kinds of research have been carried out in the modelling of the spatial distribution analysis of soil properties and heavy metals (Guo et al., 2015; Heung et al., 2016; Hengl et al., 2018; Pouladi et al., 2019; Kebonye et al., 2020; Kebonye et al., 2021). Yet, there are still limitations in different ML. For example, the widely used random forest (RF) is posed with challenge that it requires enormous computational power and resources to build numerous trees and combine their outputs. Also, RF outputs are difficult to interpret. Also, the RF limitation is similar to the cubist model as they are both tree-based models (Zhou et al., 2019).

Gaussian process regression (GPR) is new and rarely applied in soil mapping among different ML methods. Gaussian process regression is a generic supervised learning method designed to solve regression and probabilistic classification problems (Rasmussen and Williams 2006). Its prediction advantages are computing empirical confidence intervals and deciding based on whether one should refit (online fitting, adaptive fitting) the prediction in some region of interest. GPR does not need a semivariogram model, and it can accommodate several covariates and automate their implementation. It can also be combined with other interpolation methods, creating a hybrid method to improve its prediction efficiency. The merit of GPR over other machine learning techniques is that the algorithms models both the expectation and the variance of the random variable, thus permitting mapping the prediction uncertainty Ballabio et al.(2019). Also, GPR allows the specification of the input data noise, so if prior knowledge about it is known, it can be used to avoid overfitting the data. Ballabio et al.(2019) adopted GPR to map LUCAS topsoil chemical properties because of its capability to produce uncertainty maps and prediction variance. In the shallow landslide susceptibility mapping conducted by Colkesen et al. (2016), GPR outperformed logistic regression to predict landslides.

Meanwhile, Gonzalez agreed that GPR is an excellent technique in creating low-cost digital soil maps. Besides that, Kumar et al. (2012) and Mirzaee et al. (2016) reported that a hybrid model could incorporate the spatial autocorrelation of measured variables to achieve better predictions and lower errors. Similarly, Li et al. (2011) stated that hybrid models produced 30% more accurate predictions than any other method. According to Shadrin et al. (2021), their proposed hybrid model, GPR- Bayesian Information Criterion (BIC) approach, showed better performance on average with a 15% higher  $R^2$  score than other geostatistical models.

Therefore, this present study tries to show GPR and GPR hybrid model excellent performances rather than comparing them with other machine learning models. The GPR hybrid model proposed in this study is new and have not been applied before and, as such, contributes to the growing works of literature.



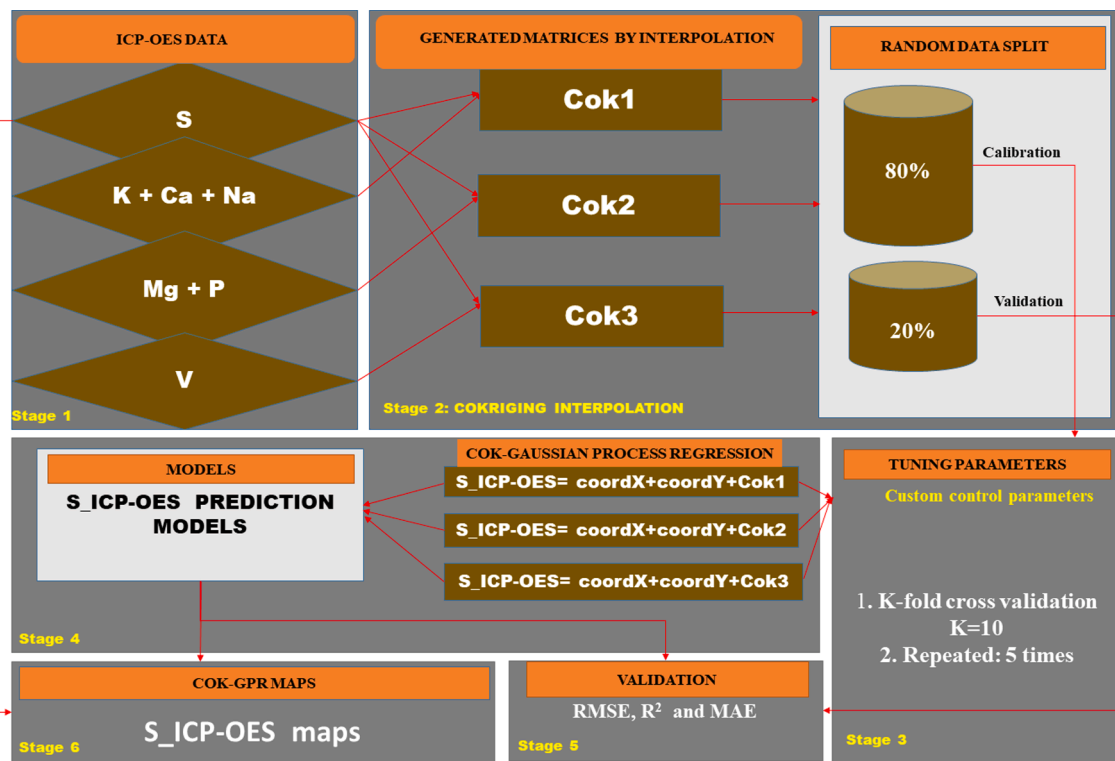


Fig. 2. Workflow of the study.

Therefore, how good is a hybrid method's performance for interpolating S levels compared to only GPR or Cok interpolation for soil S? Thus, this work's objective was to propose a hybrid method based on the cokriging (Cok) and ML technique [i.e. Gaussian process regression (GPR)]. The proposed hybrid method was compared with the Cok and GPR modelling methods.

## 2. Materials and methods

### 2.1. Research location

The research site is situated in Moravian-Silesian Region's foothills, Frýdek-Místek district in the Czech Republic (Fig. 1). It is an active agricultural site situated at geographical coordinates of latitude 49°41'0"N and longitude 18°20'0"E and the elevation of 225–327 m above sea level (Agyeman et al., 2020b). Meanwhile, according to the Koppen classification system, the area's climate is classified as Cfb = Temperate oceanic climate with a high rainfall even in dry months. The study area is approximately 889.8 km<sup>2</sup> designated for agricultural activities with scattered trees.

The regions' soils are predominantly Cambisols and occupy approximately 56.7% of the Czech Republic agricultural land (Vacek et al. 2020). They are characterized by fine-textured materials derived from a wide range of rocks, primarily colluvial and alluvial (Němeček & Kozák 2003).

### 2.2. Soil sampling

Soil samples were collected through a grid sampling technique with a stainless steel bucket auger. Composite soil samples were collected (n = 115) from the topsoil at a depth of 0–20 cm into a well-labelled Ziploc bag and transported to the Soil Science and Soil Protection Department's laboratory at the Czech University Life Sciences, Prague for analysis.

### 2.3. Laboratory studies

Soil samples were air-dried under laboratory conditions and pulverized to achieve a fine powder between 3 and 4 μm size via an automatic mill (Hanchen Soil Crusher with 220 V pattern name). Soil samples were treated with Aqua regia reagent (a mixture of HCl and HNO<sub>3</sub> in the ratio of 3:1). The mixture was used to extract the soil pseudo-total concentration of elements according to Tejnecký et al. (2015) and Cools and De Vos (2016). The pseudo-total elements were measured via Inductively Coupled Plasma Optical Emission Spectroscopy (ICP-OES) model iCAP 7000. Soil analysis with ICP-OES was performed in duplicates and later averaged, and a blank sample was also intermittently measured via ICP-OES. Furthermore, in this present study, S, Ca, K, Mg, Na, P, and V were selected and presented as S\_ICP-OES, Ca\_ICP-OES, K\_ICP-OES, Mg\_ICP-OES, Na\_ICP-OES, P\_ICP-OES, and V\_ICP-OES, respectively. These elements represent some of the essential constituents of soils.

### 2.4. Cokriging (Cok) interpolation

Cokriging employs several variable types to predict a particular target variable (in this case, soil sulphur). These variables must also exhibit a strong relationship with the targeted property (Bivand, 2008; Tziachris et al., 2019). A simple Cok equation is presented in equation (2).

$$Z'(x_0) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (1)$$

where  $Z'(x_0)$  is the predicted/interpolated value for point  $x_0$ ,  $Z(x_i)$  is the known value, and  $\lambda_i$  is the kriging weight for the  $Z(x_i)$  values. It can be calculated by the semivariance function of the variables on the condition that the estimated value is unbiased and optimal (Equation (3)).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^n [Z(X_i) - Z(X_i + h)]^2 \quad (2)$$

**Table 1**  
Cokriging matrix prediction.

Matrix X					vector y	
sampling points (n)	Coordinates X	Coordinates Y	S_ICP-OES	}	Cok <sub>i..n</sub>	
1	-464432	-1124117	132.6358			346.2833
2	-471856	-1116632	371.0277			419.1817
3	-477924	-1121975	373.2175			387.7108
4	-464186	-1117977	283.6727			550.9499
5	-462146	-1128047	262.4481			354.1073
6	-469935	-1115740	371.9971			394.9518
	"	"	"			"
	"	"	"			"
	"	"	"			"
	"	"	"			"
	"	"	"			"
	"	"	"			"
	"	"	"			"
n	-433653	-1140010	214.3023		212.8357	
n-1	-434002	-1135953	147.321		188.4218	

**Table 2**  
Summary statistics of the response and predictors variables.

	Response	Predictors					
	S_ICP-OES	Ca_ICP-OES	K_ICP-OES	Mg_ICP-OES	Na_ICP-OES	P_ICP-OES	V_ICP-OES
	mg/kg						
Mean	403.7 ± 22.0	3624.8 ± 743.2	1289.7 ± 41.7	1981.9 ± 62.2	150.7 ± 15.4	682.0 ± 33.5	31.4 ± 0.9
Standard Deviation	236.3	7969.7	446.9	666.7	165.3	358.8	9.4
Kurtosis	16.7	54.2	4.8	11.7	15.0	13.8	8.7
CV	58.5	219.9	34.6	33.6	109.6	52.6	29.8
Skewness	3.4	7.2	1.5	2.5	2.9	3.1	2.1
Minimum	131.0	538.7	497.5	685.7	7.1	294.6	15.6
Maximum	1815.4	69161.8	3535.7	5970.1	1202.9	2903.1	81.9
Confidence Level (95.0%)	43.7	1472.2	82.5	123.2	30.5	66.3	1.7

where  $\gamma(h)$  is the semivariance,  $N(h)$  is the point group number at distance  $h$ ,  $Z(x_i)$  is the numerical value at position  $x_i$ , and  $Z(x_i + h)$  is the numerical value at a distance  $(x_i + h)$ .

For Cok, we selected the best model based on how small MAE and RMSE (closer to 0) are and how large  $R^2$  (closer to 1) is.

2.5. Gaussian process regression (GPR)

The Gaussian Process (GPR) is a non-parametric modelling approach (Vasudevan et al., 2009; Wang et al., 2020; Zhang and Xu, 2020). This is generally a supervised learning method designed to solve regression and probabilistic classification problems. The current study estimated the relationship between S\_ICP-OES levels and some selected soil elements (Ca\_ICP-OES, K\_ICP-OES, Mg\_ICP-OES, Na\_ICP-OES, P\_ICP-OES, and V\_ICP-OES), respectively. The GPR is useful because of its simplicity and reasonable accuracy, credited by Wang et al. (2020). Moreover, GPR can help lower data overfitting (Ballabio et al., 2019). To describe GPR, both

the mean  $[m(x)]$  and covariance/kernel  $[k(x_i, x_j)]$  functions are used (Seeger, 2004). In expression form, this is given as:

$$f(x) \text{ GP}[m(x), k(x_i, x_j)] \tag{3}$$

The  $x$  in equation (4) represents each input vector. Mean and covariance functions can be further expressed as equations (5) and (6), separately.

$$m(x) = E[f(x)] \tag{4}$$

$$k(x_i, x_j) = \text{cov}[f(x_i), f(x_j)] \tag{5}$$

Similar to Wang et al. (2020), “The squared exponential (SE) covariance function with a discrete length scale for each predictor is used to fit the GPR model”:

$$k(x_i, x_j | \theta) = \sigma_f^2 \exp \left[ -\frac{1}{2} \sum_{m=1}^d \frac{(x_{im} - x_{jm})^2}{\sigma_m^2} \right] \tag{6}$$

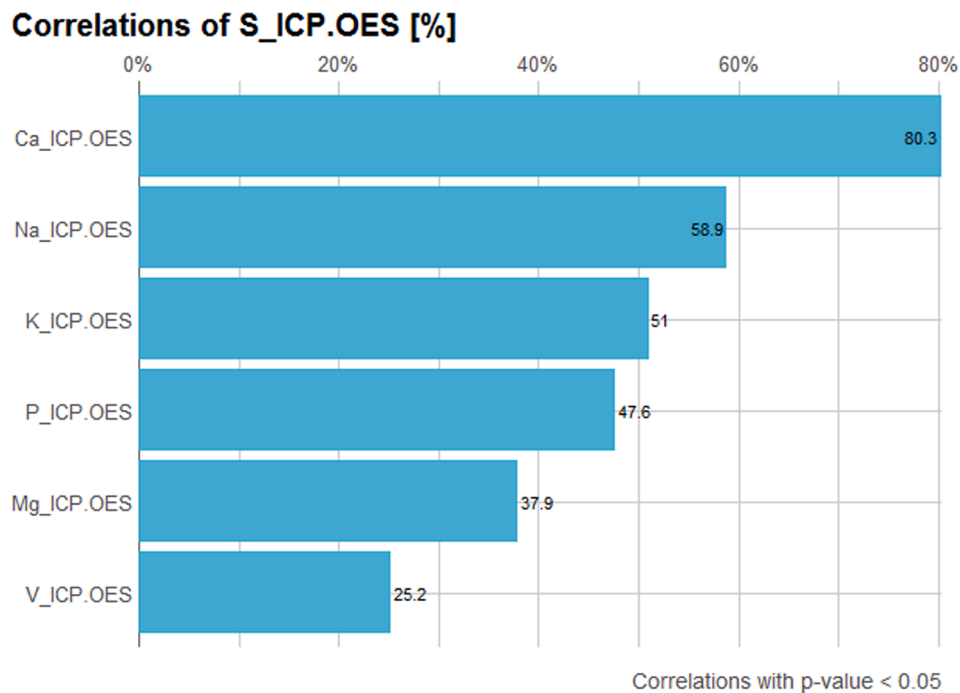


Fig. 3. Correlation matrix showing the percentage of the relationship between the response and predictors ( $p < 0.05$ ).

**Table 3**  
spatial dependency of the three models.

Models	Nugget	Sill	Range	Model fitness	Spatial dependency
Cok1	40167.19	74,539.97	3724.364	Gaussian	0.54
Cok2	34581.19	67,030.74	3724.364	Gaussian	0.51
Cok3	32719.34	66,190.23	3351.042	Gaussian	0.49

**Table 4**  
Cokriging model accuracy estimation.

Models	Cokriging prediction		
	MAE mg/kg	RMSE mg/kg	R <sup>2</sup>
Cok1	-1.28	164.42	0.85
Cok2	2.26	223.77	0.79
Cok3	1.24	241.51	0.80

The  $\sigma_m$  represents the length scale for each predictor variable  $m$  ( $m = 1, 2, \dots, d$ ),  $\sigma_f$  denotes the signal standard deviation (Zhang and Xu, 2020). The  $\theta$  is such that:

$$\theta_m = \log \sigma_m \quad (7)$$

Further explanations regarding GPR are presented in Vasudevan et al. (2009) and Ballabio et al. (2019). The 'caret' package was used with set method value `gaussprLinear` executed through the `kernelab` library in the R environment.

## 2.6. Cokriging-Gaussian process regression (Cok-GPR) interpolation

In order to establish the interpolation technique via the proposed hybrid model (Cok-GPR), a training matrix was defined as obtained via the Cok method. The workflow is presented in Fig. 2. In the X matrix, the coordinates (coordinates X and coordinates Y) of the point and the predicted value of the target soil variable obtained by using the Cok method (i.e.,  $Cok_i$ ,  $i = 1$  to  $n$ ), estimated based on the highest to a moderately significant correlation observed between S\_ICP-OES and

each of the predictors ( $p < 0.05$ , 0.01, 0.1), were considered as independent variables (features). Different X matrix (coordinates X, coordinates Y and  $Cok_i$ ,  $i = 1$  to  $n$ ) via the Cok model were obtained based on the different levels of correlation. The known value at the point itself was not considered in the calculation of the Cok attribute. Thus, the matrix generated, in which the  $n$  rows represented the  $n$  sampling points, and three columns were the independent variables (coordinates X, coordinates Y, with  $Cok_i$ ,  $i = 1$  to  $n$ ). The known value of the soil attribute at the sampled point (S\_ICP-OES) was considered the model's dependent variable ( $y$ ). Fig. 1 represents the COK-GPR method's training set divided into independent variables (matrix X) and dependent variable (vector y) (Table 1).

Furthermore, using ML for predictive mapping of targeted soil properties, georeferenced coordinates, soil properties, reflectance values obtained by satellite images, sensor data, among other data, can also be used to construct an ML model. Similarly, after obtaining X, Y and  $Cok_i$ ,  $i = 1$  to  $n$  (Table 1) via Cok, they were then applied to the model for S\_ICP-OES via the COK-GPR model. The dependent variable represents the measured soil S (S\_ICP-OES), whose values are predicted at unsampled locations in the COK-GPR model. The matrix X and the vector y were the inputs of the training set of the Cok-GPR method.

## 2.7. Design of the evaluation experiment

The proposed hybrid method was compared with CoK and GPR, respectively. For GPR and Cok-GPR, 80% of the datasets were used for calibration, while the remaining 20% was used for validation. The calibration parameters of the models were tuned using repeated k-fold cross-validation, with  $k = 10$ , to avoid overfitting.

The mean absolute error (MAE), root mean square error (RMSE) and the coefficient of determination ( $R^2$ ) was used to evaluate the model as well as map performances (John et al., 2020; Kebonye et al., 2021). For MAE and RMSE, a lower value is preferred, while for  $R^2$ , a larger value is always expected. Based on Li et al. (2016),  $R^2 \geq 0.75$  is considered a good prediction.

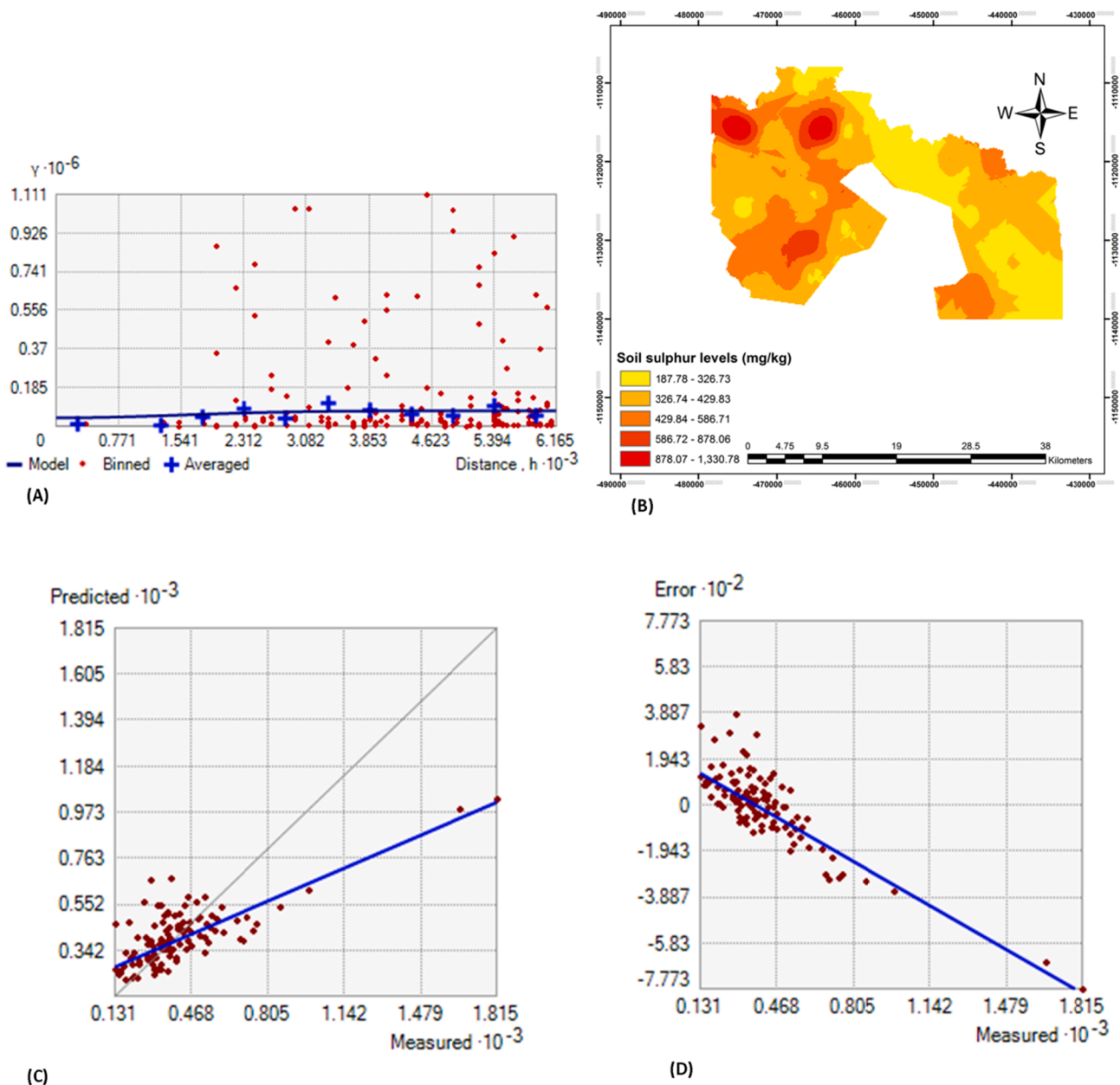


Fig. 4. (A) Semivariance fitted by Gaussian model, y-axis (semivariance) and x-axis (distance) are presented in standard forms, respectively, (B) Prediction map by Cok1, estimated via highly correlated predictors (C) Fitness curve, y-axis (predicted values in mg/kg) and x-axis (observed values in mg/kg) are expressed in standard forms, respectively (D) Error plot, y-axis (error values in mg/kg) and x-axis (observed values in mg/kg) are presented in standard forms.

### 3. Results and discussion

#### 3.1. Samples descriptive statistics

The response variable ( $S_{\text{ICP-OES}}$ ) ranged from 131.0 to 1815.4 mg/kg, with a mean of  $403.7 \pm 22.0$  mg/kg (Table 2).  $S_{\text{ICP-OES}}$  showed a strong coefficient of variation with positively skewed sample points. The soil sulphur content obtained in the present cultivated land is higher than 114–373 mg/kg reported by Kopittke et al. (2019) and 241 to 391 mg/kg by Srinivasarao et al. (2004) in Indian soils. Meanwhile, the high amount of  $S_{\text{ICP-OES}}$  obtained in the area may be attributed to the type of fertilizer and cultivation period (Srinivasarao et al., 2004; Wang et al., 2007; Kopittke et al., 2019). Kopittke et al. (2019) emphasized that increase P fertilizers and soil cultivation without adequate replenishment results in the loss of total sulphur. Therefore, the high amount of  $S_{\text{ICP-OES}}$  may be attributed to the decrease in P rich fertilizers and hard

coal deposition from the steel site nearby (Sager, 2012; Agyeman et al., 2020b). The higher S levels may also be attributed to the top reached value (484 kg S/ha) experienced in 1991 in North Bohemia (Balík et al., 2009).

Also presented in Table 2 is the summary statistics of the predictors. The means of the predictors are  $3624.8 \pm 743.2$  mg/kg,  $1289.7 \pm 41.7$  mg/kg,  $1981.9 \pm 62.2$  mg/kg,  $150.7 \pm 15.4$  mg/kg,  $682.0 \pm 33.5$  mg/kg and  $31.4 \pm 0.9$  mg/kg for Ca<sub>ICP-OES</sub>, K<sub>ICP-OES</sub>, Mg<sub>ICP-OES</sub>, Na<sub>ICP-OES</sub>, P<sub>ICP-OES</sub> and V<sub>ICP-OES</sub>, respectively. All the predictors showed a strong coefficient variation (>26%) and a positive skewness. In this study, the values obtained for Ca<sub>ICP-OES</sub> and Mg<sub>ICP-OES</sub> were generally higher than that of Jodral-Segado et al. (2006). They reported total calcium and magnesium content of  $34.09 \pm 7.80$  mg/kg and  $14.23 \pm 2.25$  mg/kg, respectively, in Spain's agricultural soil. Ca<sub>ICP-OES</sub> value obtained here was lower than of the soils of Albania (Shallari et al. 1998) but higher than of Belgium (De Temmerman et al. 2003) and

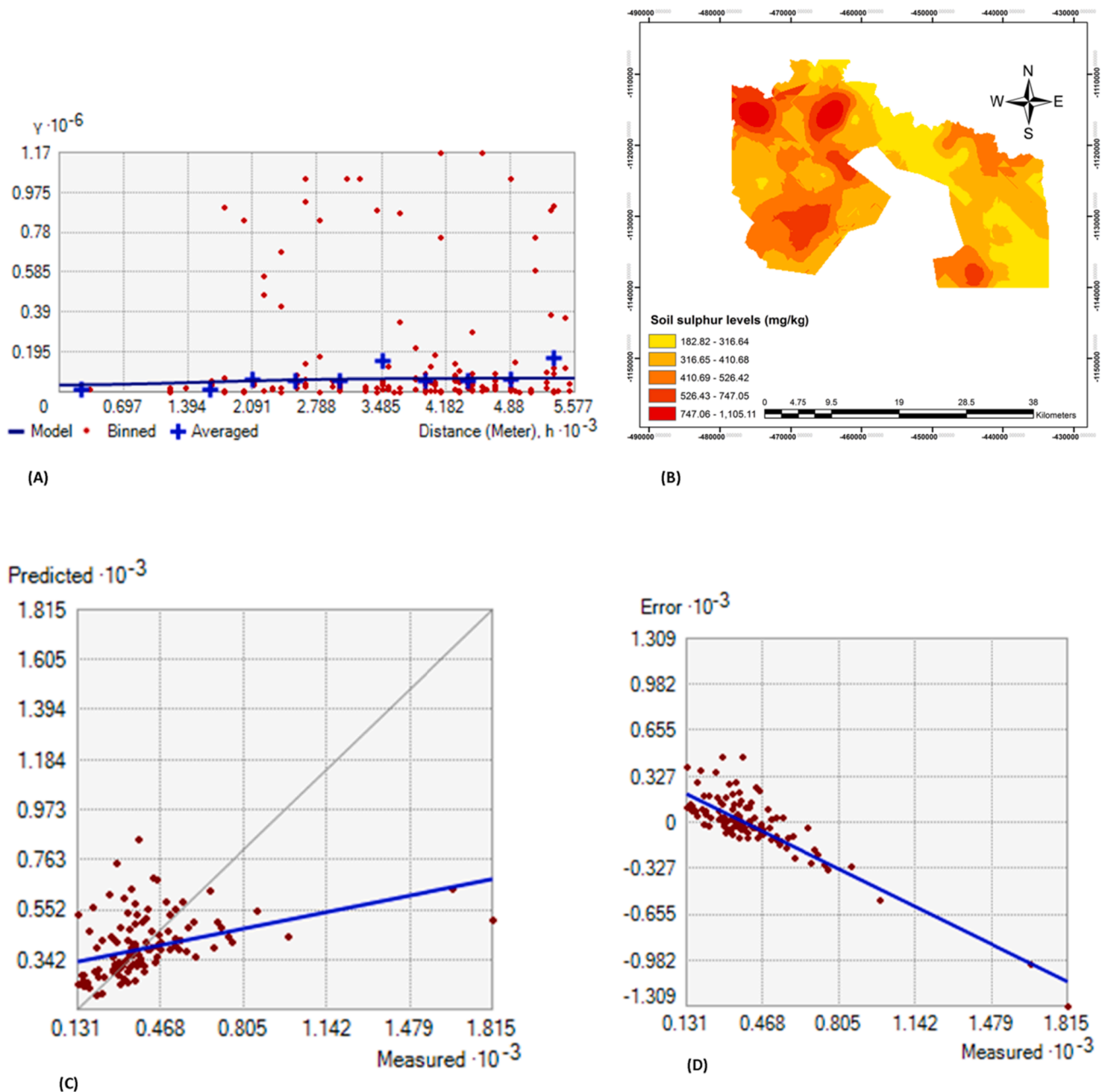


Fig. 5. (A) Semivariance fitted by Gaussian model, y-axis (semivariance) and x-axis (distance) are presented in standard forms, respectively, (B) Prediction map by Cok2, estimated via moderately correlated predictors (C) Fitness curve, y-axis (predicted values in mg/kg) and x-axis (observed values in mg/kg) are expressed in standard forms, respectively (D) Error plot, y-axis (error values in mg/kg) and x-axis (observed values in mg/kg) are presented in standard forms.

Italy (De Nicola et al. 2003). Similarly,  $K_{ICP-OES}$  was higher than extracted K of Norway soils (65–155 mg/kg) as reported by Løes and Øgaard (2003), and  $Na_{ICP-OES}$  was found to be within the recommended range of 20–250 mg/kg of the agricultural soils of Kenya (Akenga et al., 2014).  $P_{ICP-OES}$  content was way higher than the  $105 \pm 1.29$  mg/kg reported by Akenga et al. (2014) in Kenya’s soils, and 20 mg/kg threshold recommended for optimal plant growth by Li et al. (2014). On the other hand,  $V_{ICP-OES}$  was the only selected trace element observed to be lower than the average world value of 129 mg/kg reported by Kabata-Pendias (2011). The result suggested that  $V_{ICP-OES}$  did not enrich the soils of the cultivated land. In general, all the predictors were within optimal cropping levels, as indicated by Imran et al. (2010). The decreasing sequence abundance of predictors in the studied soils, as observed in Table 1 was  $Ca_{ICP-OES} > Mg_{ICP-OES} > K_{ICP-OES} > P_{ICP-OES} > Na_{ICP-OES} > V_{ICP-OES}$ .

### 3.2. Correlation of sulphur with other variables

The result of the correlation between  $S_{ICP-OES}$  and the predictors is presented in Fig. 3. The correlation was significant at  $p < 0.05$  with the predictors being ranked, showing how strong, moderate, weak or very weak the relationship between them and the response variable is. This result was fundamental in building our Cok model, which is then piped into the COK-GPR model. All the relationships between  $S_{ICP-OES}$  and the predictors were positive. However,  $S_{ICP-OES}$  showed a strong correlation with  $Ca_{ICP-OES}$  ( $r = 80.3\%$ ,  $p < 0.05$ ), a moderate correlation with  $K_{ICP-OES}$  ( $r = 58.9\%$ ,  $p < 0.05$ ) and  $Na_{ICP-OES}$  ( $r = 51\%$ ,  $p < 0.05$ ), a weak correlation with  $P_{ICP-OES}$  ( $r = 47.6\%$ ,  $p < 0.05$ ) and  $Mg_{ICP-OES}$  ( $37.9\%$ ,  $p < 0.05$ ) and then a very weak correlation with  $V_{ICP-OES}$  ( $r = 25.2\%$ ,  $p < 0.05$ ).

$Ca_{ICP-OES}$  strong correlation output with  $S_{ICP-OES}$  showed a

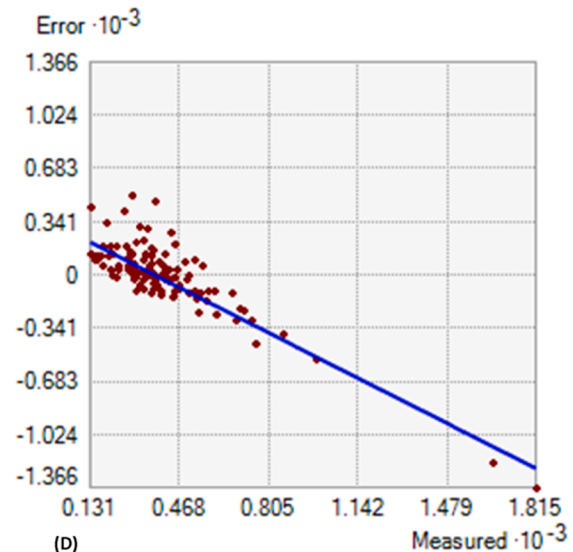
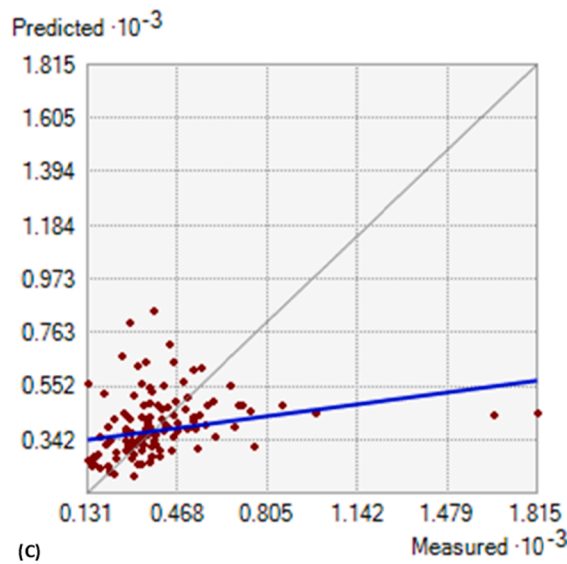
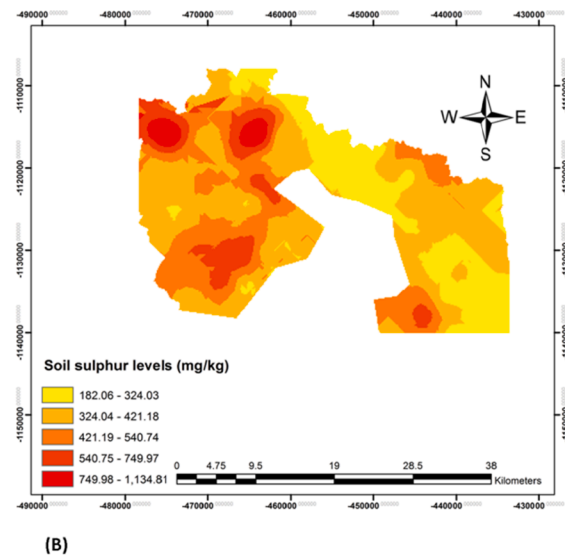
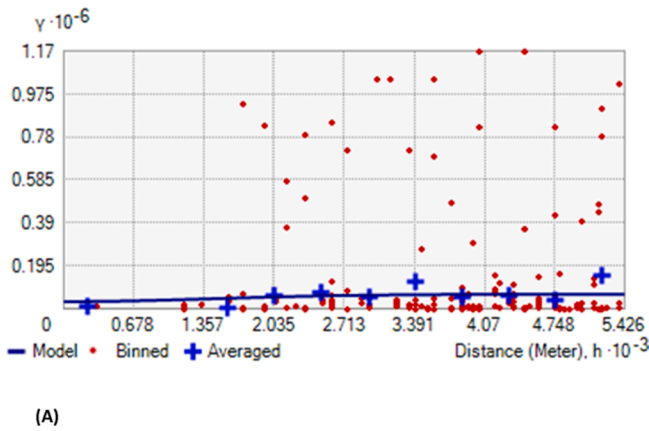


Fig. 6. (A) Semivariance fitted by Gaussian model, y-axis (semivariance) and x-axis (distance) are presented in standard forms, (B) Prediction map by Cok3, estimated via weakly correlated predictors (C) Fitness curve, y-axis (predicted values in mg/kg) and x-axis (observed values in mg/kg) are expressed in standard forms, (D) Error plot, y-axis (error values in mg/kg) and x-axis (observed values in mg/kg) are presented in standard forms.

**Table 5**  
Comparison of model prediction accuracy via GPR.

Models	Gaussian process regression prediction					
	Calibration			Validation		
	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>
	mg/kg	mg/kg		mg/kg	mg/kg	
GPR1	77.50	108.09	0.67	85.43	137.59	0.83
GPR2	171.20	171.20	0.31	180.50	303.51	0.18
GPR3	119.30	171.86	0.23	187.87	326.30	0.05

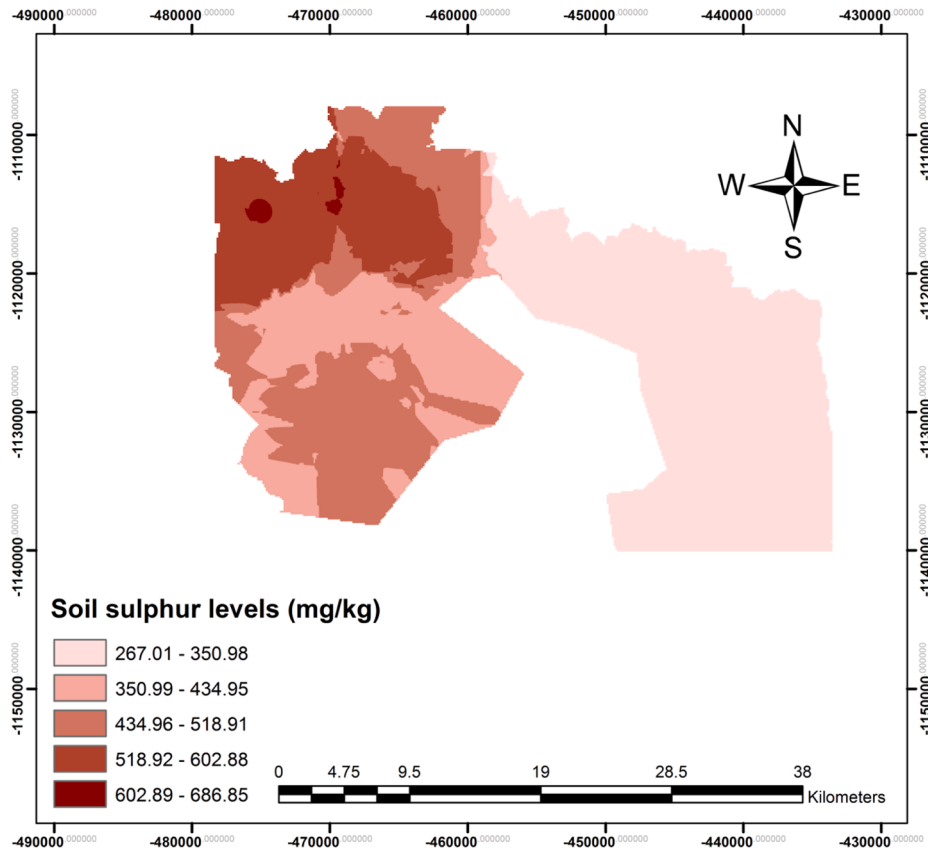
strong linear relationship between them as reported by Aulakh and Dev (1976) and Srinivasarao et al. (2004). Deficiency in Ca<sub>ICP-OES</sub> could result in S<sub>ICP-OES</sub> deficiency and vice versa; however, they are principal gypsum constituents (Argani, 1968). The moderate correlation output obtained between S<sub>ICP-OES</sub> and K<sub>ICP-OES</sub>, and Na<sub>ICP-OES</sub>, respectively, was supported by Saha et al. (2013). They observed a significant increase in sulphur when soils were amended by potassium

fertilizer. In the case of V<sub>ICP-OES</sub>, the output is supported by the works of Zhang et al. (2018). They stated that S<sub>ICP-OES</sub> could serve as bio-oxidation in the removal of V<sub>ICP-OES</sub>.

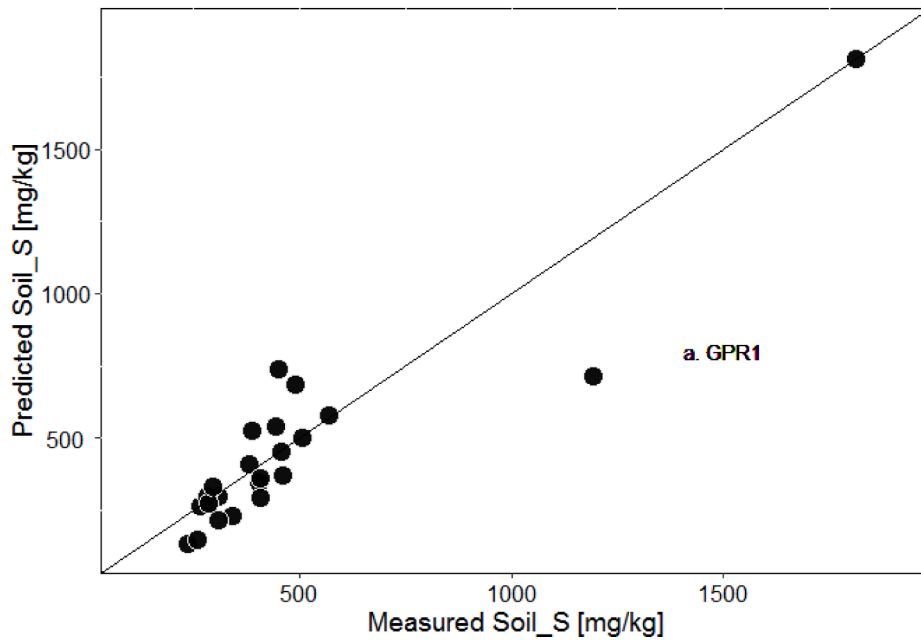
Furthermore, the considered nutrient elements (except for K<sub>ICP-OES</sub>), Ca<sub>ICP-OES</sub>, Mg<sub>ICP-OES</sub>, and Na<sub>ICP-OES</sub> are generally absorbed in lesser amounts than P<sub>ICP-OES</sub> and S<sub>ICP-OES</sub>, but much more than the V<sub>ICP-OES</sub>. Mg<sub>ICP-OES</sub> deficiency is more common than Ca<sub>ICP-OES</sub> deficiency but much less common than K<sub>ICP-OES</sub>. Na<sub>ICP-OES</sub> is not essential to plant growth. Agricultural soils are tested for Na<sub>ICP-OES</sub> to diagnose sodic, saline-sodic problems and potential irrigation challenge (Di Meo et al., 2003).

### 3.3. Cokriging modelling

In Fig. 2, S<sub>ICP-OES</sub> was interpolated via the highly correlated predictors (Ca<sub>ICP-OES</sub>, Na<sub>ICP-OES</sub> and K<sub>ICP-OES</sub>), represented as Cok1. S<sub>ICP-OES</sub> interpolation was also performed via moderately correlated predictors (P<sub>ICP-OES</sub> and Mg<sub>ICP-OES</sub>), expressed as Cok2 and then via

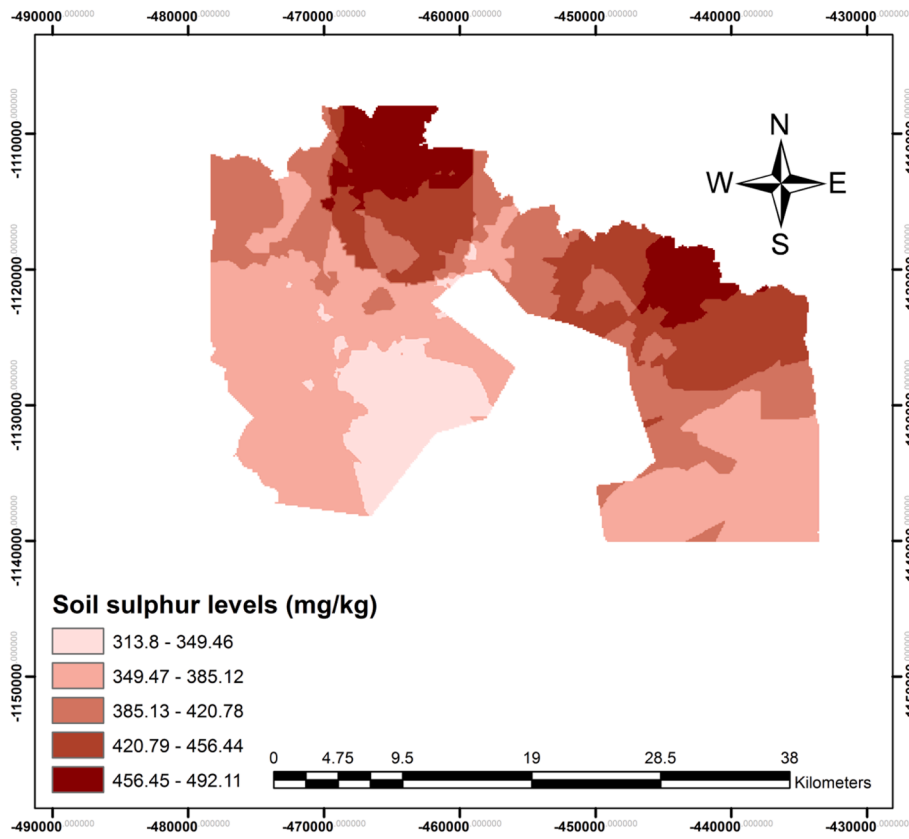


(A) GPR2 map

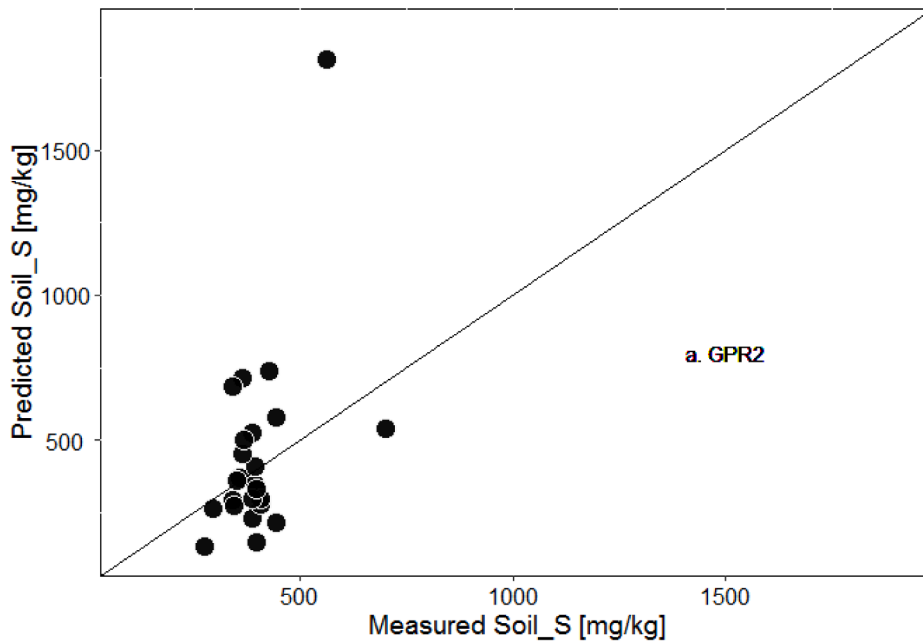


(B) Fitness curve

Fig. 7. GPR1 prediction map and fitness curve.



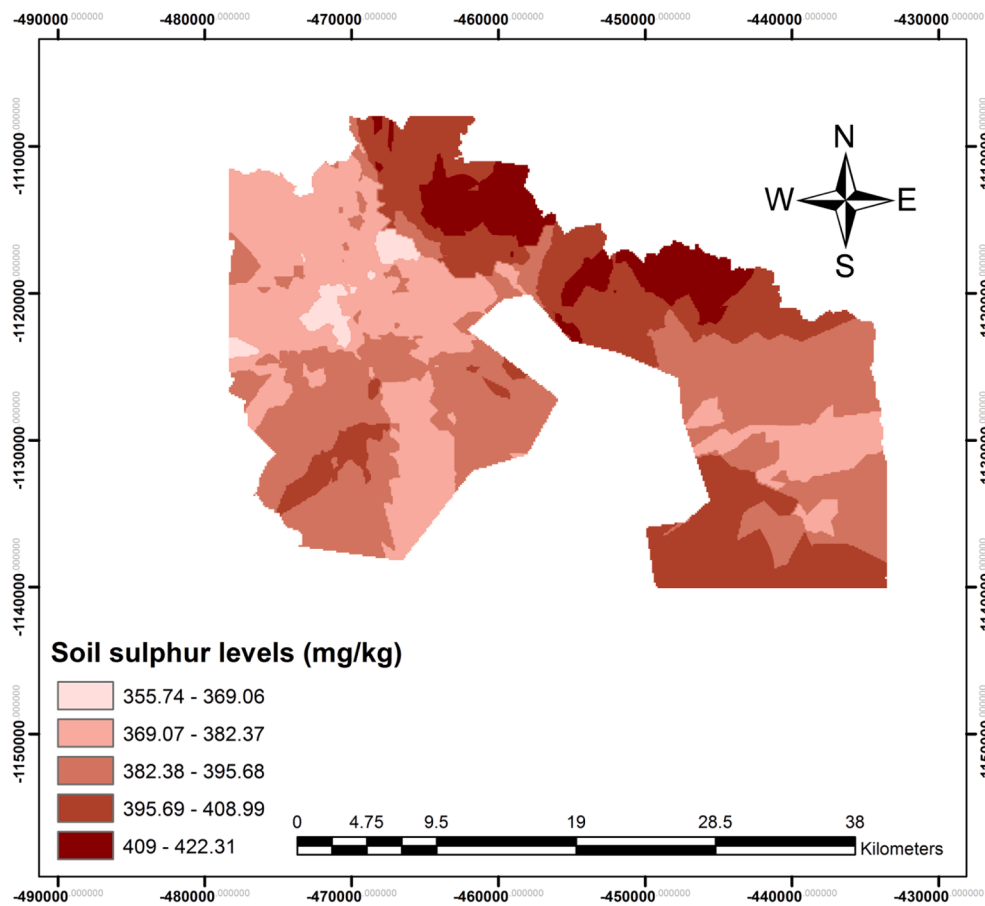
(A) GPR2 map



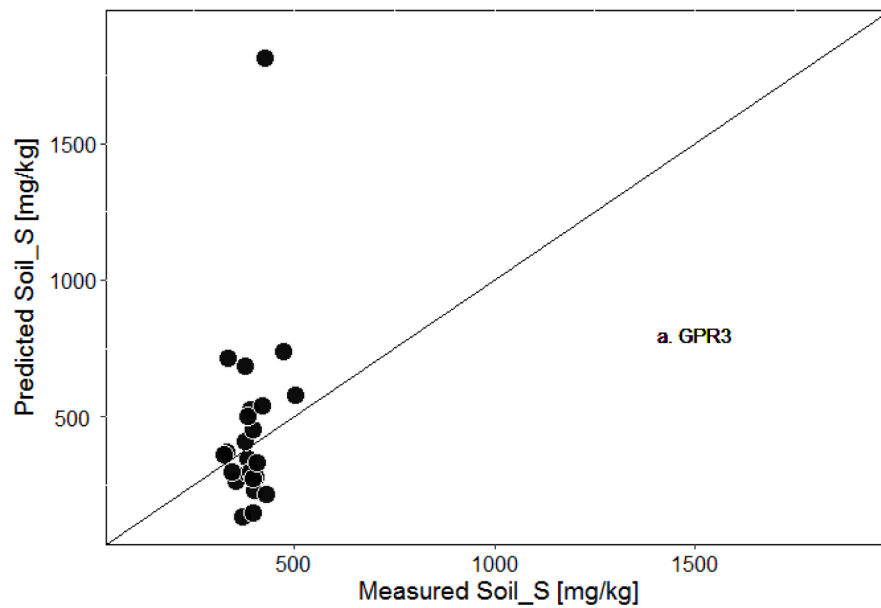
(B) Fitness curve

Fig. 8. GPR2 prediction map and fitness curve.





(A) GPR3 map



(B) Fitness curve

Fig. 9. GPR3 prediction map and fitness curve.

**Table 6**  
Comparison of model prediction accuracy via COK-GPR modelling.

Models	Cokriging-Gaussian process prediction					
	Calibration			Validation		
	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>
	mg/kg	mg/kg		mg/kg	mg/kg	
Cok1-GPR	62.22	79.83	0.78	76.84	102.11	0.91
Cok2-GPR	70.24	94.45	0.75	87.65	138.35	0.84
Cok3-GPR	68.75	93.69	0.78	83.31	126.18	0.87

the weakly correlated predictor (V<sub>ICP-OES</sub>), described as Cok3. This approach was supported by Yao et al. (2012) and Shi et al. (2012). Cok1, Cok2 and Cok3 were all fitted by a Gussain model and yielded a moderate spatial dependency (SPD) ( $0.25 < SPD \leq 0.75$ ) (Table 3).

Furthermore, the result of the cokriging model accuracy estimation is presented in Table 4. All the data were implemented to assess the quantitative mapping accuracy of S<sub>ICP-OES</sub>. Cok1, Cok2 and Cok3 prediction outputs were compared using MAE, RMSE and R<sup>2</sup> criteria. An MAE value close to zero indicates a lack of bias, RMSE should be as small as possible, and R<sup>2</sup> should close to 1. Compared to Cok1 and Cok2 predictions, Cok1 had the smallest ME and RMSE, which indicates a better prediction of S<sub>ICP-OES</sub> is achieved via highly significantly correlated auxiliary variables. Cok2 had a smaller RMSE than Cok3, while Cok3 had a smaller MAE than Cok2. The R<sup>2</sup> (>0.75) values were within the acceptable predictions for Cok1, Cok2 and Cok3, respectively. However, the overall best performing model is Cok1, which applied highly significantly correlated auxiliary variables. The result obtained here is similar to the report by Song et al. (2014).

The semivariogram, maps, and fitness curves results are presented in Figs. 4, 5 and 6. Sulphur in the present soils exhibited high spatial variability. The result obtained showed that the soils in the western region of the study area had high contents of S (>1000 mg/kg) when the prediction was done using the three approaches (Cok1, Cok2 and Cok3). However, S<sub>ICP-OES</sub> contents in the eastern part of the study area were low ( $\leq 421$  mg/kg) except for the South-east and northeast, where S<sub>ICP-OES</sub>'s values were  $410 \text{ mg/kg} < S_{ICP-OES} \leq 540 \text{ mg/kg}$ . Nevertheless, the values obtained for the study do not indicate any S<sub>ICP-OES</sub> deficiency. The result obtained from the present investigation is comparable to the report of Olson and Englestad (1972), where Mollisols (S<sub>ICP-OES</sub> = 500 mg/kg), Alfisols (S<sub>ICP-OES</sub> = 400 mg/kg) had values similar to those reported in this study. However, a Eutrudand from Hawaii (S<sub>ICP-OES</sub> = 1280 mg/kg) recorded higher S<sub>ICP-OES</sub> than those reported in this study's eastern soils.

### 3.4. Gaussian process regression (GPR) modelling

The comparison of the three models' results is presented in Table 5, while the interpolated maps and fitness curve are presented in Figs. 6, 7 and 8, respectively. During calibration, GPR1 presented an accepted calibrated model output (MAE = 77.50 mg/kg, RMSE = 108.09 mg/kg and R<sup>2</sup> = 0.67) while GPR2 and GPR3, gave a poor calibration. In the validation process, GPR1 dominated as the best performing model (MAE = 85.43 mg/kg, RMSE = 137.59 mg/kg and R<sup>2</sup> = 0.83). Generally, all the models were far from having a perfect regression value (R<sup>2</sup> = 0.05 – 0.83 relative to R<sup>2</sup> = 1.00) and the dissimilarities between the models are evident from the visual assessment (Figs. 7–9). GPR1 (using Ca<sub>ICP-OES</sub>, Na<sub>ICP-OES</sub> and K<sub>ICP-OES</sub> as predictors) was completely different from GPR2 (using P<sub>ICP-OES</sub> and Mg<sub>ICP-OES</sub> as predictors). Overall, GPR1 models (using Ca<sub>ICP-OES</sub>, Na<sub>ICP-OES</sub> and K<sub>ICP-OES</sub> predictors) could better predict total soil S content in the study area. GPR has shown to be advantageous as it can handle hidden non-linear relationships between variables, confirming why it might have performed better than each of the GPR2 and GPR3.

Sulphur prediction using Gaussian process regression showed narrower ranges of S<sub>ICP-OES</sub> compared with cokriging. Sulphur's high

values mainly were observed in the western part of the study area, while low prediction values were observed in the eastern part of the study area (<350 mg/kg) when the prediction was made using GPR1. However, utilizing GPR2 showed high S values in the Northern and North-eastern parts of the study area. Low values mainly were observed in the western and South-eastern parts of the study area. When the GPR3 approach was used, only soils around the northern and southern part of the study area had high S values, and other regions had low values.

### 3.5. Cokriging-Gaussian process regression (Cok-GPR) modelling

Presented in Table 6 is the accuracy and comparison of Cok-GPR modelling and maps presented in Figs. 10–12. The tuning parameters of the dataset incorporated the k = 10-fold cross-validation, repeated five times. Cok1-GPR yield a good calibration model (MAE = 62.22 mg/kg, RMSE = 79.83 mg/kg and R<sup>2</sup> = 0.78) compared to Cok2-GPR and Cok3-GPR, respectively. In the model validation, Cok1-GPR still stood out as the best model (MAE = 76.84 mg/kg, RMSE = 102.11 mg/kg and R<sup>2</sup> = 0.91) when compared with Cok2-GPR and Cok3-GPR. The hybrid model (Cok-GPR) showed an improvement in the S<sub>ICP-OES</sub> prediction in the agricultural soils of Frydek Mistek. The difference presented by Cok-GPR models when compared to GPR is obvious. For example, MAE and RMSE were reduced by a greater percentage by reducing the variation between the observed values and the predicted values. Generally, all the Cok-GPR models were within the range of good models (R<sup>2</sup> > 0.75). Like Gaussian process regression, S prediction using Cok-GPR showed narrower ranges of S<sub>ICP-OES</sub> contrary to Cok. S<sub>ICP-OES</sub>'s high values were mostly observed in the western part of the study area, while low prediction values were observed in the eastern part of the study area (<386 mg/kg) when the prediction was performed using Cok1. However, utilizing Cok2 and Cok3, the result showed high S values around the West, South-east and Northern part of the study. Low values were mostly observed in the eastern part of the study area. Also, to make up for uncertainty in the prediction maps of the hybrid model, the standard deviation maps are presented in the supplement list (S1). Furthermore, this predictive method has not been applied before, and the result obtained here cannot be compared with any results. However, the high performance of the hybrid model followed a similar output pattern obtained in the GPR-BIC model by Shadrin et al. (2021). Cok1-GPR, Cok2-GPR, and Cok3-GPR produced approximately R<sup>2</sup> of 7%, 6%, and 9% higher than Cok1, Cok2 and Cok3.

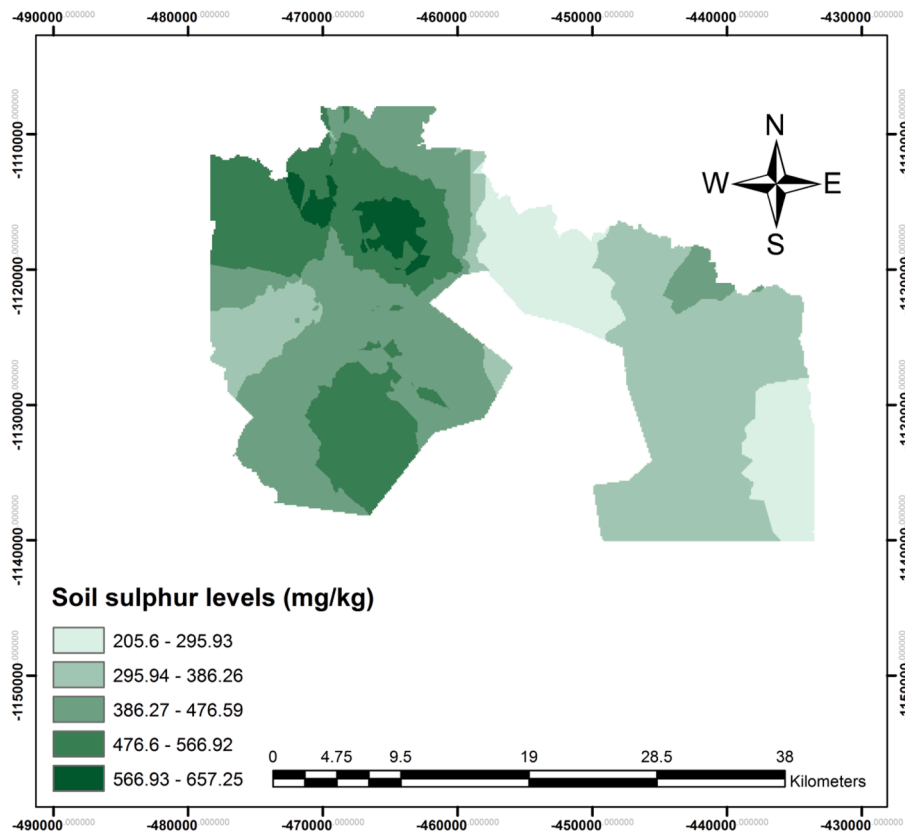
### 3.6. Selecting the best model

We further evaluated the performance of all the models via the Taylor diagram using the software provided by Agrisoft (Taylor, 2005). Fig. 13 shows the Taylor diagram plot. The result revealed that all the models produced similar normalized standard deviation values (i.e., between 0.75 and 1). However, the ratio (i.e. ratio of the model standard deviation to the reference value standard deviation) and the model accuracy value (Table 7) presented the differences in the models. Overall, our hybrid model (Cok-GPR) produced the best performing model.

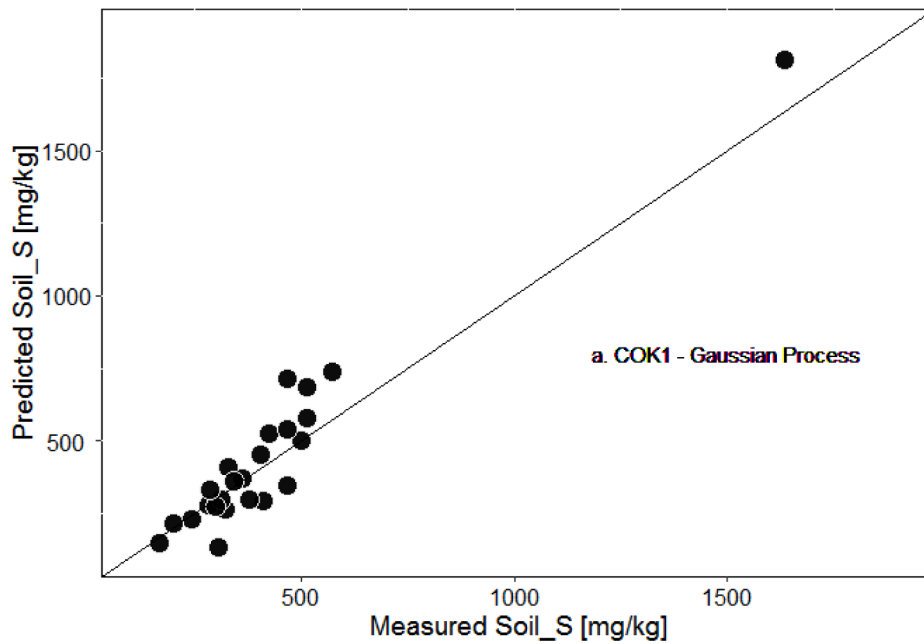
## 4. Conclusion

In conclusion, in this presented study, different covariates that showed a significant correlation with soil S were used for modelling. This approach was seen to improve the prediction of soil S via GPR using Ca, Na and K as predictors.

Furthermore, in this work, the hybrid model (Cok-GPR), which combines cokriging and Gaussian process regression, improved both Cok and GPR models' fitting accuracy, respectively. The model was able to naturally take care of uncertainty that might have been originated from the cokriging models. Cok-GPR was also able to handle the unevenly spaced but correlated training datasets provided by the different Cok (i.e. Cok1, Cok2 and Cok3) models.

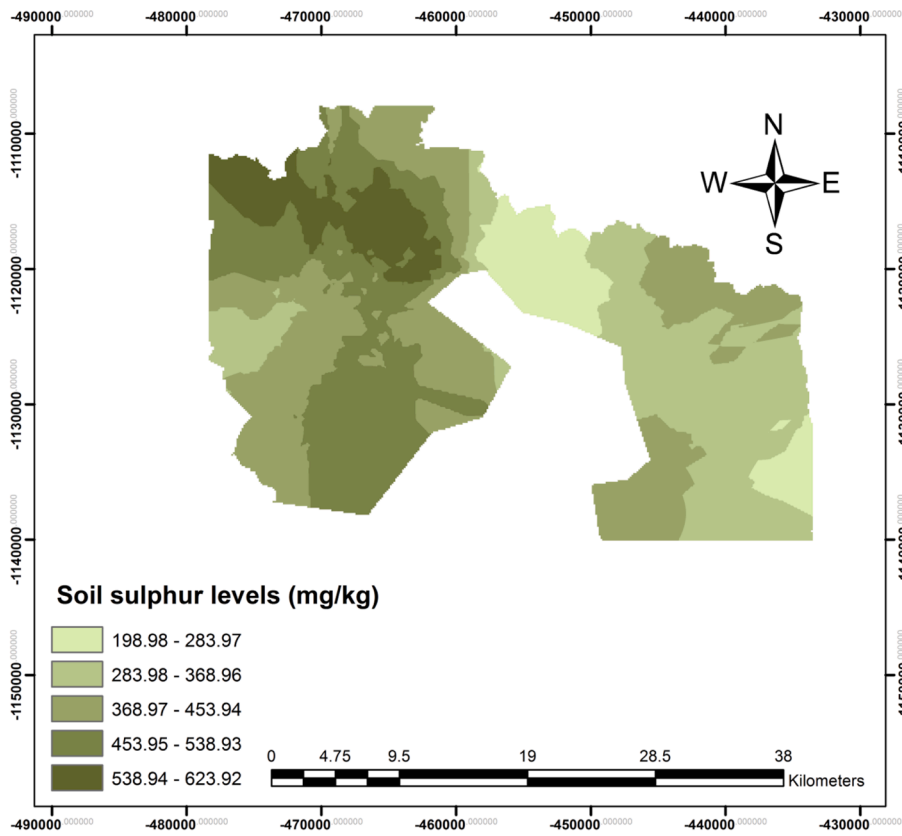


(A) Cok1-GPR map

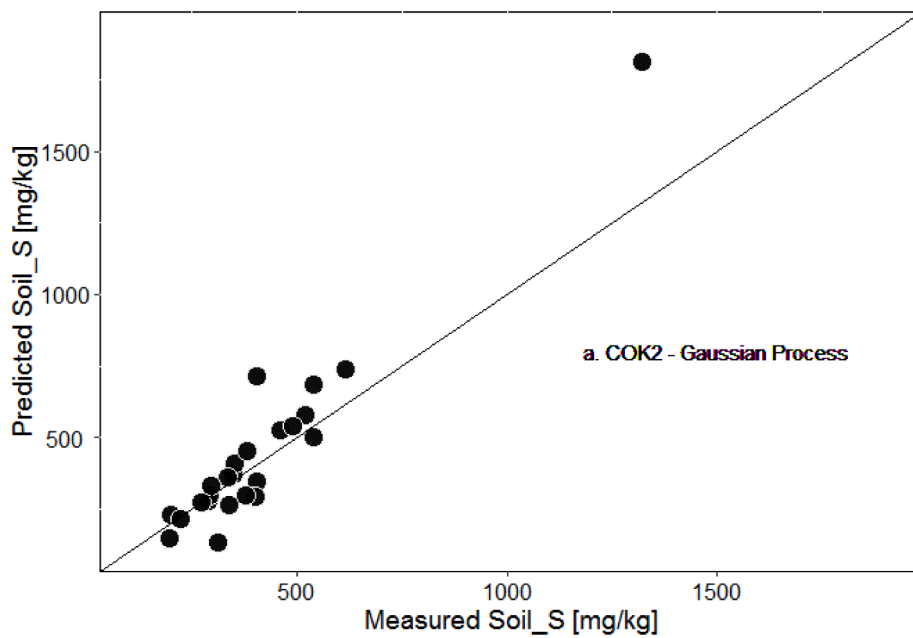


(B) Fitness curve

Fig. 10. Cok1-GPR prediction map and fitness curve.

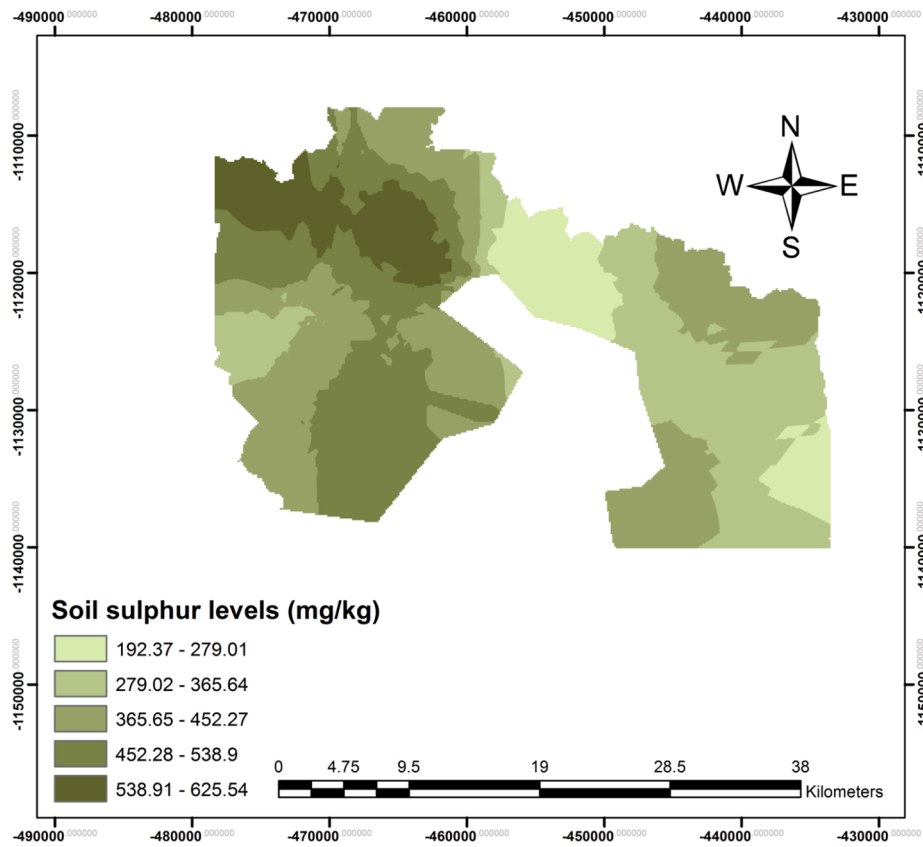


(A) Cok2-GPR map

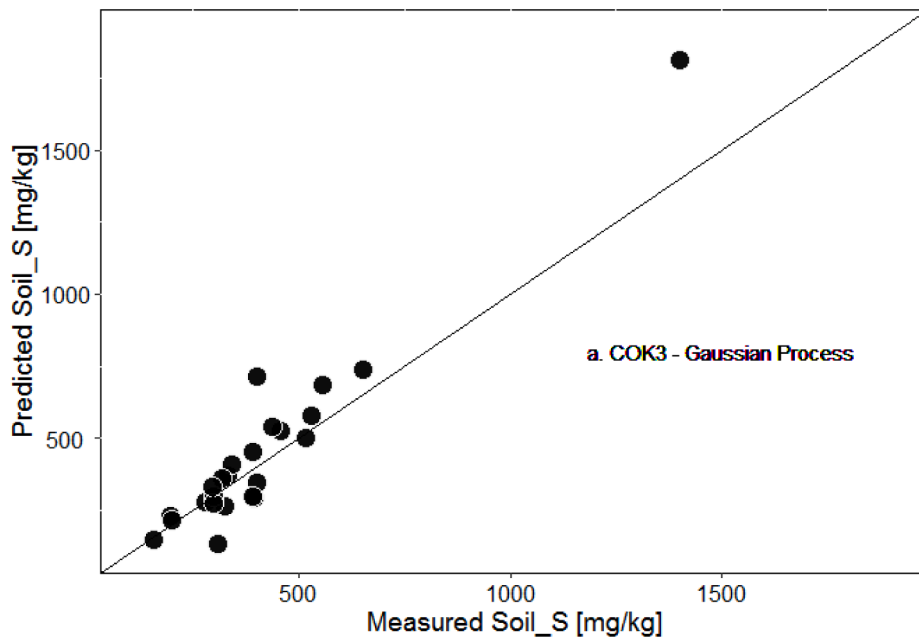


(B) Fitness curve

Fig. 11. Cok2-GPR prediction map and fitness curve.

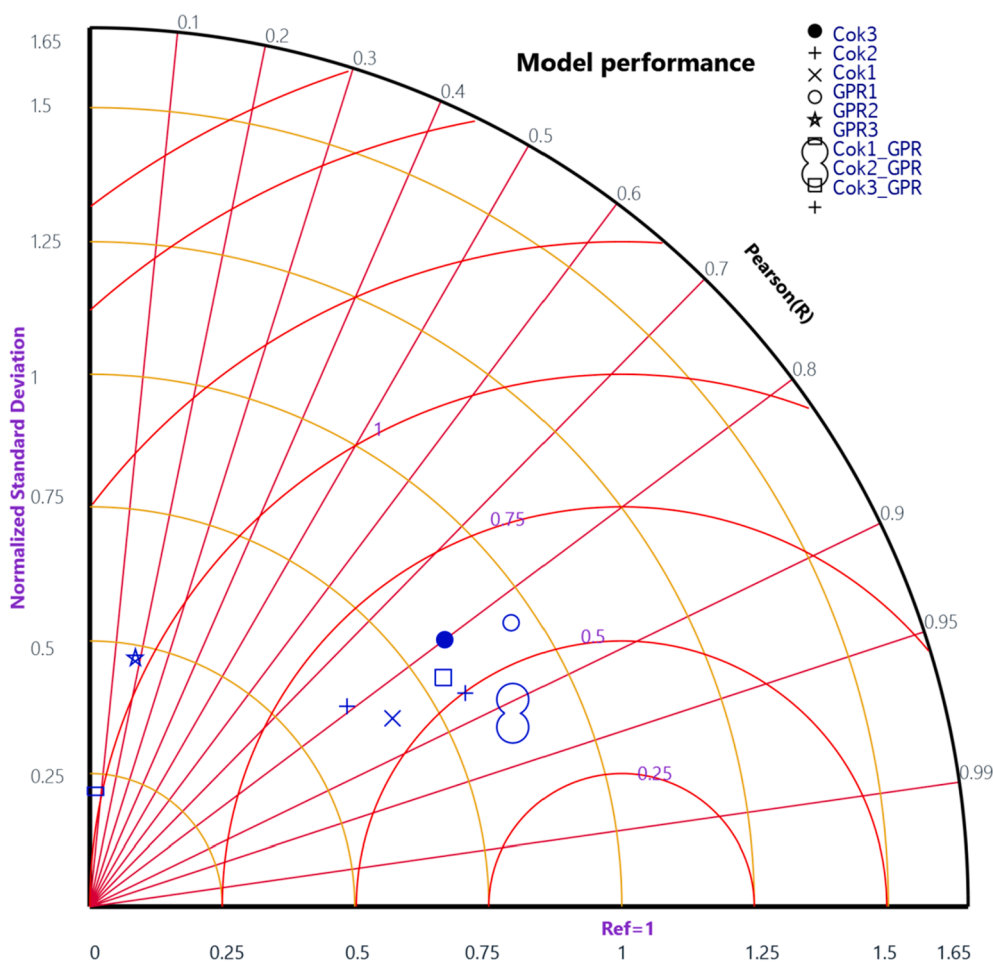


(A) Cok3-GPR map



(B) Fitness curve

Fig. 12. Cok3-GPR prediction map and fitness curve.



**Fig. 13.** A Taylor Diagram was revealing the best performing model. The red coloured semi-circles on the x-axis represent root mean square (RMSE) values. The  $-90^\circ$  curves travelling from the y-axis to the x-axis represent the standard deviation (STD) values. The brick-red coloured straight line emanating from the origin is the Pearson correlation value, and Ref = 1 is the reference value.

**Table 7**  
Performance of all the models in predicting soil sulphur.

Name	Ratio value	Coefficient of determination
Cok3	0.834	0.80
Cok2	0.612	0.79
Cok1	0.669	0.85
GPR1	0.954	0.83
GPR2	0.476	0.18
GPR3	0.216	0.05
Cok1_GPR	<b>0.874</b>	<b>0.91</b>
Cok2_GPR	0.791	0.84
Cok3_GPR	<b>0.811</b>	<b>0.87</b>

The study demonstrated that the proposed Cok-GPR model was able to show more precisely the soil S levels spatial distribution in the actively cultivated agricultural soil. The models showed that the Cok-GPR model had higher fitting accuracy and robustness than COK and GPR models. Nevertheless, Cok-GPR computational cost is significantly higher. Moreover, Cok1-GPR (using coordinate x, coordinate y and predicted cokriging soil Sulphur matrix as predictors) provided the best model. The proposed Cok-GPR model can potentially be applied to predict soil nutrient element levels efficiently, do proper soil fertilization calculations, and apply to do more precise soil management practices.

**Funding**

This study was supported by an internal PhD grant no. SV20-5-21130 of the Faculty of Agrobiology, Food and Natural Resources of the Czech University of Life Sciences Prague (CZU). And also, NutRisk grant: European Regional Development Fund, project Center for the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially harmful substances of anthropogenic origin: comprehensive assessment of soil contamination risks for the quality of agricultural products, number CZ.02.1.01/0.0/0.0/16\_019/0000845.

**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.

**Acknowledgment**

The authors are grateful to the Staff of the Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food, and Natural Resources, Czech University of Life Sciences, Prague, Professor Luboš Borůvka and Radim Vašát for guiding in the sampling design; Samuel K Ahado for his assistance in processing the samples and Ondřej Drábek for his expertise in the laboratory protocol. The support from the

Ministry of Education, Youth and Sports of the Czech Republic (project No. CZ.02.1.01/0.0/0.0/16\_019/0000845) is also acknowledged.

## References

- Argani, L., 1968. Effect of gypsum and its principal constituents (Ca and S) on yield of hay and seed of lucerne. *Sementic Ellette* 14, 170–177.
- Agyeman, P.C., Ahado, S.K., Borůvka, L., Biney, J.K.M., Sarkodie, V.Y.O., Kebonye, N.M., Kingsley, J., 2020a. Trend analysis of global usage of digital soil mapping models in the prediction of potentially toxic elements in soil/sediments: a bibliometric review. *Environ. Geochem. Health* 1–25.
- Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., Kocarek, M., 2020b. Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. *Environ. Geochem. Health* 1–20.
- Akenga, P., Salim, A., Onditi, A., Yusuf, A., Waudou, W., 2014. Determination of selected micro and macronutrients in sugarcane growing soils at Kakamega North District, Kenya. *IOSR J. Appl. Chem* 7 (7), 4–41.
- Aulakh, M.S., Dev, G., 1976. Profile distribution of sulphur in some soil series of Sangrur district, Punjab. *J. Indian Soc. Soil Sci.* 24, 308–313.
- Balík, J., Kulhánek, M., Černý, J., Száková, J., Pavlíková, D., Čermák, P., 2009. Differences in soil sulfur fractions due to limitation of atmospheric deposition. *Plant Soil Environ.*
- Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. *Geoderma* 355, 113912.
- Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2008. *Applied Spatial Data Analysis with R* Vol. 747248717, 237–268.
- Colkesen, I., Sahin, E.K., Kavzoglu, T., 2016. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. *J. African Earth Sci.* 118, 53–64.
- Coelho, A.L.F., Queiroz, D.M., Valente, D.S.M., Pinto, F. de A. de C., 2018. An open source spatial analysis system for embedded systems. *Comput. Electron. Agric.* 154, 289–295.
- Cools, N., De Vos, B., 2016. Part X: Sampling and analysis of soils. In: *UNECE ICP Forests Programme Coordinating Centre (Ed.): Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effect of air pollution on forests*. Thünen Institute of Forest Ecosystems, Eberswalde, Germany.
- De Nicola, F., Maisto, G., Alfani, A., 2003. Assessment of nutritional status and trace element contamination of Holm oak woodlands through analyses of leaves and surroundings soil. *Sci. Total Environ.* 311, 191–203.
- De Temmerman, L., Vanongelal, L., Boon, W., Hoening, M., Geypens, M., 2003. Heavy metal content of arable soils in Northern Belgium. *Water Air Soil Pollut.* 148, 61–76.
- Di Meo, V., Michele, A., Paola, A., Pietro, V., 2003. Availability of potassium, calcium, magnesium, and sodium in “bulk” and “rhizosphere” soil of field grown corn determined by electro ultrafiltration. *J. Plant Nutr.* 26 (6), 1149–1168.
- Drake, J.M., Randin, C., Guisan, A., 2006. Modelling ecological niches with support vector machines. *J. Appl. Ecol.* 43, 424–432.
- Fox, R.L., Hasan, S.M., Jones, R.C., 1971. Phosphate and sulfate sorption by Latosols. In: *Proceedings International Symposium on Soil Fertility Evaluation (New Delhi)*, pp. 857–864.
- Giacomin, G., Carvalho, M.B., Santos, A. de P., Medeiros, N. das G., Ferraz, A.S., 2014. Comparative analysis of interpolation methods for surface models. *Rev. Bras. Cartogr.* 66, 1315–1329.
- Gautam, R., Panigrahi, S., Franzen, D., Sims, A., 2011. Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. *Biosyst. Eng.* 110, 20–28.
- Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma* 237, 49–59.
- Hedge, N.G., Mujumdar, S., P R, R.N., Jambarmath, S.S., R P, M., 2017. Survey paper on agriculture yield prediction tool using machine learning. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* 5, 36–39.
- Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 120, 75–93.
- 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, e5518.
- Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* 265, 62–77.
- Imran, M., Kanwal, F., Livi, M., Amir, M., Iqbal, M.A., 2010. Evaluation of physico-chemical characteristics of soil samples collected from Harrapa-Sahiwal (Pakistan). *Asian J. Chem.* 22, 4823.
- John, K., Isong, I.A., Kebonye, N.M., Ayito, E.O., Agyeman, P.C., Afu, S.M., 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* 9, 487.
- Jodral-Segado, A.M., Navarro-Alarcón, M., De La Serrana, H.L.G., Lopez-Martinez, M.C., 2006. Calcium and magnesium levels in agricultural soil and sewage sludge in an industrial area from Southeastern Spain: relationship with plant (*Saccharum officinarum*) disposition. *Soil and Sediment Contamination: An International Journal* 15 (4), 367–377.
- Kabata-Pendias, A., 2011. *Trace Elements in Soils and Plants*. CRC Taylor and Francis Group, London New York.
- Kebonye, N.M., Eze, P.N., John, K., Gholizadeh, A., Dajčl, J., Drábek, O., Nemeček, K., Borůvka, L., 2020. Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils. *J. Geochem. Explor.* 106680.
- Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Nemeček, K., Drábek, O., Borůvka, L., 2021. Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma* 384, 114792.
- Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. *Appl. Math. Model.* 81, 401–418.
- Kopittke, P.M., Menzies, N.W., Wang, P., McKenna, B.A., Lombi, E., 2019. Soil and the intensification of agriculture for global food security. *Environ. Int.* 132, 105078.
- Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* 189, 627–634.
- Nemeček, J., Kozák, J., 2003. Approaches to the solution of a soil map of the Czech Republic at the scale 1:250 000 using SOTER methodology. *Plant Soil Environ.* 49 (7), 291–297.
- Pei, T., Qin, C.-Z., Zhu, A.-X., Yang, L., Luo, M., Li, B., Zhou, C., 2010. Mapping soil organic matter using the topographic wetness index: a comparative study based on different flow-direction algorithms and kriging methods. *Ecol. Indic.* 10 (3), 610–619.
- Li, J., Heap, A., Potter, A., Daniell, J.J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. *Environ. Model. Softw.* 26 (12), 1647–1659.
- Li, M., Yang, X.W., Tian, X.H., Wang, S.X., Chen, Y.L., 2014. Effect of nitrogen fertilizer and foliar zinc application at different growth stages on zinc translocation and utilization efficiency in winter wheat. *Cereal Res. Commun.* 42 (1), 81–90.
- Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Cong, R., Ren, T., 2016. Methods for estimating leaf nitrogen concentration of winter oilseed rape (*Brassica napus L.*) using in situ leaf spectroscopy. *Ind. Crops Prod.* 91, 194–204.
- Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. *Sensors (Switzerland)* 18, 1–29.
- Løes, A.K., Øgaard, A.F., 2003. Concentrations of soil potassium after long-term organic dairy production. *Int. J. Agric. Sustain.* 1 (1), 14–29.
- Lucheta, A.R., Lambais, M.R., 2012. Sulfur in agriculture. *R. Bras. Ci. Solo* 36, 1369–1379.
- McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. *Geoderma* 117 (1–2), 3–52.
- McClung, A.C., de Freitas, L.M.M., Lott, W.L., 1959. Analysis of several Brazilian soils in relation to plant responses to sulfur. *Soil Sci. Soc. Am. Proc.* 23, 221–224.
- Mirzaee, S., Ghorbani-Dashtaki, S., Mohammadi, J., Asadi, H., Asadzadeh, F., 2016. Spatial variability of soil organic matter using remote sensing data. *Catena* 145, 118–127.
- Olson, R.A., Englestad, O.P., 1972. *Soil phosphorus and sulfur*. Soils of the Humid Tropics. National Academy of Sciences, Washington, DC, pp. 82–101. Parton WJ, RL Sanford.
- Parmley, K.A., Higgins, R.H., Ganapathysubramanian, B., Sarkar, S., Singh, A.K., 2019. Machine learning approach for prescriptive plant breeding. *Sci. Rep.* 9, 17132.
- Pouladi, N., Möller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma* 342, 85–89.
- Qin, C., An, Y., Liang, P., Zhu, A., Yang, L., 2021. Soil property mapping by combining spatial distance information into Soil Land Inference Model (SoLIM). *Pedosphere* 31 (4), 638–644.
- Rasmussen, C.E., Williams, C.K.I., 2006. *Gaussian processes for machine learning*.
- Sager, M., 2012. Levels of sulfur as an essential nutrient element in the soil-crop-food system in Austria. *Agriculture* 2 (1), 1–11.
- Saha, S., Saha, M., Saha, A.R., Mitra, S., Sarkar, S.K., Ghorai, A.K., Tripathi, M.K., 2013. Interaction effect of potassium and sulfur fertilization on productivity and mineral nutrition of sunhemp. *J. Plant Nutr.* 36 (8), 1191–1200.
- Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. *Prog. Phys. Geog.* 27 (2), 171–197.
- Seeger, M., 2004. Gaussian processes for machine learning. *Int. J. Neural Syst.* 14 (02), 69–106.
- Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. *Remote Sens.* 12, 1–29.
- Setiyoko, A., Arymurthy, A.M., Basaruddin, T., 2019. DEM fusion concept based on the LS-SVM cokriging method. *Int. J. Image Data Fusion* 10 (4), 244–262.
- Shadrin, D., Nikitin, A., Tregubova, P., Terekhova, V., Jana, R., Matveev, S., Pukalchik, M., 2021. An automated approach to groundwater quality monitoring—geospatial mapping based on combined application of gaussian process regression and bayesian information criterion. *Water* 13 (4), 400.
- Shallari, S., Schwartz, C., Hasko, A., Morell, J.L., 1998. Heavy metals in soils and plants of serpentine and industrial sites of Albania. *Sci. Total Environ.* 209, 133–142.
- Shi, W., Liu, J., Du, Z., Yue, T., 2012. Development of a surface modeling method for mapping soil properties. *J. Geogr. Sci.* 22, 752–760.
- Song, G., Zhang, J., Wang, K., 2014. Selection of optimal auxiliary soil nutrient variables for cokriging interpolation. *PLoS ONE* 9, e99695.
- Srinivasarao, C., Ganeshamurthy, A.N., Ali, M., Singh, R.N., Singh, K.K., 2004. Sulphur fractions, distribution, and their relationships with soil properties in different soil types of major pulse-growing regions of India. *Communications in soil science and plant analysis* 35 (19–20), 2757–2769.
- Stevenson, F.J., 1986. *Cycles of Soil*. John Wiley and Sons, New York.
- Taylor Diagram Primer, 2005. Karl E. Taylor. January 2005.

- Tejnecký, V., Šamonil, P., Grygar, T.M., Vašát, R., Ash, C., et al., 2015. Transformation of iron forms during pedogenesis after tree uprooting in a natural beech-dominated forest. *CATENA* 132, 12–20.
- Tziachris, P., Aschonitis, V., Chatzistathis, T., Papadopoulou, M., 2019. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. *Catena* 174, 206–216.
- Vacek, O., Vašát, R., Geoderma, L.B., 2020. Quantifying the pedodiversity-elevation relations. *Geoderma* 373, 114441.
- Valente, D.S.M., Queiroz, D.M., Pinto, F. de A. de C., Santos, N.T., Santos, F.L., 2012. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. *Sci. Agric.* 69, 173–179.
- Vasudevan, S., Ramos, F., Nettleton, E., Durrant-Whyte, H., 2009. Gaussian process modeling of large-scale terrain. *J. Field Rob.* 26 (10), 812–840.
- Wang, S., Zhu, L., Fuh, J.Y.H., Zhang, H., Yan, W., 2020. Multi-physics modeling and Gaussian process regression analysis of cladding track geometry for direct energy deposition. *Opt. Lasers Eng.* 127, 105950.
- Wang, X.B., Hoogmoed, W.B., Cai, D.X., Perdok, U.D., Oenema, O., 2007. Crop residue, manure and fertilizer in dryland maize under reduced tillage in Northern China: II Nutrient balances and soil fertility. *Nutr. Cycling Agroecosys.* 79, 17–34.
- Webster, R., Oliver, M., 2001. *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.
- Webster, R., Oliver, M.A., 1992. Sample adequately to estimate variograms of soil properties. *J. Soil Sci.* 43, 177–192.
- Yao, Y., Gao, B., Zhang, M., Inyang, M., Zimmerman, A.R., 2012. Effect of biochar amendment on sorption and leaching of nitrate, ammonium, and phosphate in a sandy soil. *Chemosphere* 89 (11), 1467–1471.
- Zhang, Y., Xu, X., 2020. Fe-based superconducting transition temperature modeling through Gaussian process regression. *J. Low Temp. Phys.* 1–14.
- Zhang, B., Qiu, R., Lu, L., Chen, X., He, C., Lu, J., Ren, Z.J., 2018. Autotrophic vanadium (V) bioreduction in groundwater by elemental sulfur and zerovalent iron. *Environ. Sci. Technol.* 52 (13), 7434–7442.
- Zhu, A.X., Band, L., Vertessy, R., Dutton, B., 1997. Derivation of soil properties using a soil land inference model (SoLIM). *Soil Sci. Soc. Am. J.* 61, 523–533.
- Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., Armaghani, D.J., 2019. Random forests and cubist algorithms for predicting shear strengths of rockfill materials. *Appl. Sci.* 9 (8), 1621.





# Estimating Soil Organic Matter: A Case Study of Soil Physical Properties for Environment-Related Issues in Southeast Nigeria

Kokei Ikpi Ofem<sup>1</sup> · Kingsley John<sup>2</sup> · Mark Pawlett<sup>3</sup> · Michael Otu Eyong<sup>1</sup> · Chukwuebuka Edwin Awaogu<sup>4</sup> · Pascal Umeugokwe<sup>4</sup> · Gare Ambrose-Igho<sup>5</sup> · Peter Ikemefuna Ezeaku<sup>4</sup> · Charles Livinus Anija Asadu<sup>4</sup>

Received: 15 August 2021 / Revised: 2 October 2021 / Accepted: 4 October 2021 / Published online: 17 October 2021  
© King Abdulaziz University and Springer Nature Switzerland AG 2021

## Abstract

The different deposition periods in sedimentary geological environment have made the build-up and estimation of soil organic matter ambiguous to study. Soil organic matter has received global attention in the ambience of international policy regarding environmental health and safety. This research was to understand the inter-relationship between soil organic matter and bulk density, saturated hydraulic conductivity (Ksat), total, air-filled and capillary porosities for organic matter estimation, via different multiple linear regression functions (i.e., *leapbackward*, *leap forward*, *leapseq* and *lmStepAIC*), in soils developed over the sedimentary geological environment. Eight mapping units were obtained in Ishibori, Agoi Ibami and Mfamosing via digital elevation model. Two pits were sited within each mapping unit, and 53 soil samples were used for the study. In soils over shale–limestone–sandstone, two pits were sited, six in alluvium, four in sandstone–limestone and four in limestone. Overall correlation between SOM with Ksat ( $r=0.626$ ) and BD ( $r=-0.588$ ) was significant ( $p<0.001$ ). The strongest correlation was obtained for SOM with BD ( $r=-0.783$ ) and Ksat ( $r=0.790$ ) in soils over limestone. In contrast, soils over shale–limestone and sandstone geological environment gave the weakest relationship ( $r<0.6$ ). Linear regression gave a similar prediction output. The best performing was *leapbackward* (RMSE = 11.50%,  $R^2=0.58$ , MAE = 8.48%), which produced a smaller error when compared with *leap forward*, *leapseq* and *lmStepAIC* functions in organic matter estimation. Therefore, we recommend applying leapback linear regression when estimating soil organic variation with physical soil properties for solving soil–environmental issues towards sustainable crop production in southeast Nigeria.

**Keywords** Agriculture · Environment · Multivariate statistics · Soil health · Humid tropics

## 1 Introduction

Soil organic matter (SOM) is an essential component of the soil. It is pivotal for maintaining multiple soil-derived ecosystem services, such as the production of food and materials

for shelter, fuel and clothing, the maintenance of biodiversity, and critically mitigating effects of global climate change (Li et al. 2017). In addition, it positively impacts soil fertility. It contains an unknown number of compounds derived from living and non-living organic substances, varying from easily decomposable simple organic materials to complex recalcitrant compounds and organisms (Kogel-Knabner 2002).

Besides sequestering or acting as a source or sink of atmospheric carbon, SOM storage in arable soils influences soil physical, chemical and biological properties (Saint-Laurent et al. 2017; Blanco-Canqui et al. 2013). These properties are exposed to more risks in cultivated soils. Land degradation occurs globally due to poor land management strategies, such as inappropriate land uses like bush burning, continuous cultivation and tillage (Blanco-Canqui et al. 2013). This results in a decline in SOM and concurrent impacts on soil physical parameters

✉ Kingsley John  
johnk@af.czu.cz

<sup>1</sup> Department of Soil Science, University of Calabar, Calabar, Nigeria

<sup>2</sup> Department of Soil Science and Soil Protection, Czech University of Life Sciences, Prague, Czechia

<sup>3</sup> School of Agrifood and Environment, Cranfield University, Bedfordshire, UK

<sup>4</sup> Department of Soil Science, University of Nigeria, Nsukka, Nigeria

<sup>5</sup> Department of Geoscience, University of Nevada-Las Vegas, Las Vegas, NV, USA

such as porosity, increased bulk density (BD) (Tisdall and Oades 1982) and reduced infiltration (Li et al. 2007) as they are functions of SOM (Jiao et al. 2020). Organic matter reduces soil BD and increases the porosity of compacted soil layers (Boni et al. 1994; Bonini and Alves, 2010), while its mineralization may lead to increased BD (Oliveira et al. 2018). In addition, researchers in Northern Karakoram (Ali et al. 2017) and Nepal (Ghimire et al. 2018) identified negative correlations between organic C and BD. In contrast and surprisingly, Lei et al. (2019) reported positive correlations between soil organic carbon (SOC) and BD in subsurface soils, while Masri and Ryan (2006) reported decreasing hydraulic conductivity with reduced SOM.

Several factors have been reported to affect the build-up of SOM. They include topography (Cardinael et al. 2017), climate (Munoz-Rojas et al. 2017), soil type (Zhao et al. 2016), soil depth, land use (Kafle, 2019), texture (Lei et al. 2019), soil microorganisms (Komarov et al. 2017) and soil pH (Zhou et al. 2020). When wholly considered, these factors make studies related to SOC complex and make its measurement and inter-relationship with other soil properties difficult. However, SOM is important in soil studies and maybe a sole indicator of fertile and healthy soil.

There have been several studies on the horizontal spatial distribution of SOM using various mathematical models as influenced by topography, vegetation and land use (Takata et al. 2007; Liu et al. 2015). Applying different machine learning (ML) in predicting soil properties is recent in soil science and precision agriculture (for example, random forest, support vector machine, artificial neural network and others) (John et al. 2020). Multiple linear regression (MLR) has been applied in modeling and predicting SOC via environmental variables and soil nutrient indicators (John et al. 2020), arsenic estimation via XRF and ICP-OES data (Kebonye et al. 2020), and the mapping of soils of Minas Gerais, Brazil via XRF data using the stepwise multiple linear regression techniques (Silva et al. 2017). However, the stepwise variable selection is automatic and has many statistical problems that could worsen if the covariates are collinear. Therefore, this study attempts to reduce covariates collinearity. Currently, no published studies compare the different stepwise linear regression functions in the modeling of SOM under diverse sedimentary geological environments; hence, this research introduces a new approach in explaining the variability of SOM in soils over the different sedimentary geological environments. We hypothesize that SOM will vary in its inter-relationship with soil physical properties in different sedimentary geological environments, and subsequently, SOM can be predicted by soil physical properties. Consequently, this research studied the inter-relationships between SOM and BD, saturated hydraulic conductivity (Ksat) and porosity, and applied various multiple

linear regression functions to predict SOM accumulation via some selected soil physical properties dominating the different geological environments.

## 2 Materials and Methods

### 2.1 Location and Land Use, Geology, and Climate of the Study Area

The study sites were located in Ishibori area (679 ha) of Ogoja (06°39'17" N, 08°47'51" E), Agoi Ibami (280 ha) in Yakurr (05°43'27"N, 08°10'37.2" E) and Mfamosing (2202 ha) in Akamkpa (05°04'41.8"N, 08° 27'49.8"E), all in the Cross River State of Nigeria. The Ogoja area is covered by the southern guinea savannah and cultivated to oil palm, teak and paddy rice, while the Yakurr and Akamkpa areas are covered by tropical rainforest. Common crops in the Yakurr and Akamkpa areas are oil palm, cassava and plantain.

Basement Complexes and Sedimentary Basins dominate the geology of Cross River State (Ekwueme 1987). The Sedimentary Basins, containing sediment fill of Cretaceous to Tertiary ages, dominate the Niger Delta region (Fatoye and Gideon 2013), with alluvium found in the low lying coastal areas. The limestone of the Cretaceous and Tertiary ages is often intercalated with shale, siltstone, and fine-grained sandstone (Ofem et al. 2020a).

Cross River State has a humid tropical climate, which varies from the southern guinea savannah in the Ogoja area to the tropical rainforest of Yakurr and Akamkpa. Consequently, rainfall fluctuates from 1251–3348 mm/year in the Ogoja area to 1760–2684 mm/year and 2109–3771 mm/year in Yakurr and Akamkpa, respectively (Sambo et al. 2016). Temperature varies from 23 to 34 °C in the Ogoja area and 23 to 32 °C in Yakurr and Akamkpa areas (Sambo et al. 2016). Yakurr and Akamkpa have similar climates and vegetation and often experience slight temperature variation.

### 2.2 Field and Laboratory Procedures

Digital elevation models (DEM) of the study locations were acquired from USGS Explorer SRTM 1 arc-second Global at a resolution of 30 m. The DEM was employed to generate slope maps in ArcGIS (ESRI, US) environment. The elevation ranges created in the slope maps were used to delineate slope transition (Ofem et al. 2020a). Each of the eight slope transitions (IH1, IH2, AI1, AI2, AI3, MF1, MF2, MF3) represented a soil mapping unit (MU). Two soil pits were randomly sited in each MU and dug to represent the soils (2 m by 1.5 m by X m). Where X m represents variable depth to the water table or consolidated rock layer, this gave rise to sixteen pits in total, two in shale–limestone–sandstone (SLM) (IH1P1, IH1P2), six

in alluvium (IH2P1, IH2P2, AI3P1, AI3P2, MF3P1, MF3P2), four in sandstone–limestone (AI1P1, AI1P2, MF1P2, MF2P2) and four in limestone (AI2P1, AI2P2, MF1P1, MF2P1). Thereby, a total of 53 soil samples were collected from pedogenic horizons and subjected to laboratory analyses. In addition, undisturbed core soil samples were vertically collected from pedogenic horizons for the determination of saturated hydraulic conductivity (Ksat), total porosity (Total\_P), air-filled porosity (Air\_P) and capillary porosity (CAP\_P). Ksat was determined by the direct application of Darcy's equation to a saturated soil column of uniform cross-sectional area (SSS, 2014), such that:

$$K_{sat} = \frac{VL}{At(H_2 - H_1)}, \quad (1)$$

where  $V$  = volume of water that flows through the sample of cross-sectional area ( $A$ ) in time ( $t$ );

$(H_2 - H_1)$  = Hydraulic head difference;

$L$  = Length of sample.

Core soil samples were then drained at 60 cm of tension to determine Total\_P, Air\_P, and CAP\_P. Total porosity, Air\_P, and CAP\_P were determined by dividing the volume of water in the soil at saturation, the volume of water drained at 60 cm of tension, and the volume of water retained at 60 cm of tension by the volume of the cylinder (Obi 2000).

Soil for organic carbon determination was air dried under room temperature in the laboratory at 29–30 °C for three days, ground with a wooden pestle to break peds and passed through a 2 mm sieve. Soil organic carbon (Walkley–Black modified acid-dichromate) was determined using standard procedures outlined in Soil Survey Staff (SSS 2014). SOM was calculated from SOC by multiplying by a factor of 1.72 to obtain SOM. Soil samples were analyzed in the Department of Soil Science, University of Nigeria, Nsukka. The field study was carried out between December 2018 and February 2019.

### 2.3 Correlation Matrix

A simple correlation analysis was performed with categorical data (e.g., geological environment). This analysis explained the intra- and inter-relationships between the SOM and the selected physical properties and how the individual geological environment contribute to the relationship between SOM and the physical properties. The output of the correlation was reconfirmed through the application of a Principal Component Analysis (PCA).

### 2.4 Principal Components Analysis

PCA enabled the grouping of the selected soil properties into the different geological environments. It enabled the extraction of principal factors accounting for the sources

of variation in the data (Belkhiri and Narany 2015) and to identify the geological material influencing SOM and other properties. Such litho-material would require further assessment as they may help explain certain SOM variability relating to the selected soil properties within the area.

### 2.5 Modeling Approach of SOM

Four ( $n=4$ ) stepwise multiple linear regression (MLR) functions were applied in this study. The forward, backward, both direction, and the regsubsets are available in the *leap* function.

This study presented four functions available in R software for stepwise linear regression in estimating SOM using six predictors (BD, Ksat, Total\_P, Air\_P, CAP, geological material). The stepwise regression applied leaps and *stepAIC* functions available in R's leaps and MASS packages. The leaps package in R is composed of "*leapBackward*", which fits a linear regression with backward selection, and "*leapForward*", with fittings for linear regression with forward selection. The "*leapSeq*" fits a linear regression with stepwise selection, while in *stepAIC* (also referred to as direction), we applied the "*lmStepAIC*" (James et al. 2014). The approach was adopted to exhaustively establish that the intended selected model is suitable for SOM prediction in the soils overlying sedimentary geological environment in the region.

The simple linear model used to predict SOM (%) via the selected soil properties is expressed as, thus:

$$SOM(\%) = \beta_0 + \sum_{j=1}^p X_j \beta_j + \epsilon_j, \quad (2)$$

where  $\beta_0$  is the  $y$ -intercept and or bias in the field of machine learning (Hastie et al. 2008). The  $X_j$  represents the predictor variable, while  $\beta_j$  is the slope coefficient of the predictor. An error term is also included and is denoted by  $\epsilon_j$ .

### 2.6 Model Accuracy and Assessment

The entire data were subjected to modeling. Mean absolute error (MAE), and root mean square error (RMSE), and coefficient of determination ( $R^2$ ) were adopted as criteria in evaluating the models' performance. In the case of MAE and RMSE, a lower value is preferred. For  $R^2$ , values closer to 1 (Li et al. 2016),

### 2.7 Statistical Analysis

The R software performed all statistical analyses and model computations (R Core Team 2019).

### 3 Results and Discussion

The summary of descriptive statistics for the soils, grouped by the geological environment, is presented in Table 1. At the same time, the results of the interaction between SOM and physical properties are shown in Fig. 1.

#### 3.1 Inter-relationships Between SOM and BD, Ksat and Porosity in the Sedimentary Geological Environment

##### 3.1.1 Soil Organic Matter Versus Bulk Density

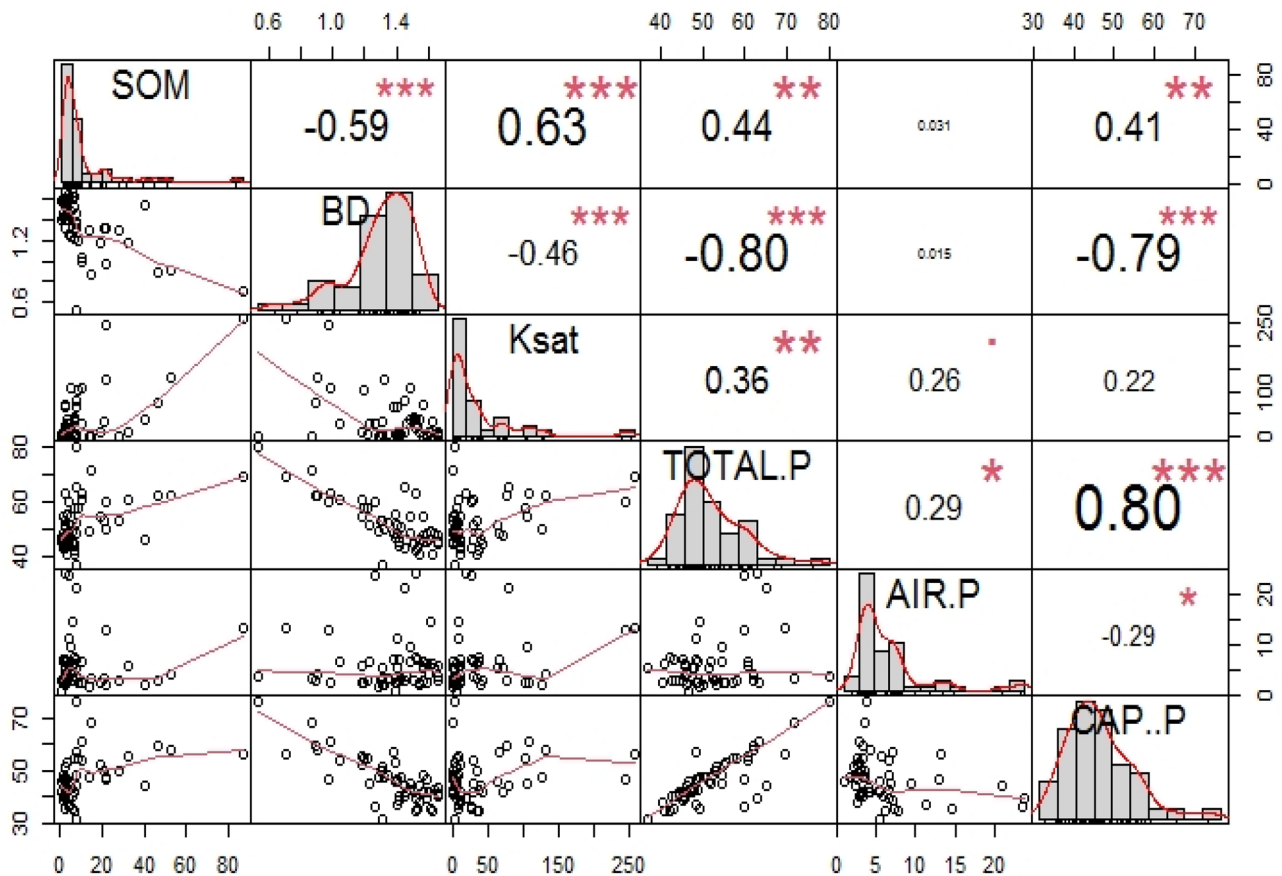
SOM correlated moderately and negatively with BD ( $r = -0.588$ ,  $p < 0.01$ ) (Fig. 1) in the studied soils and indicated an increase in SOM with decreasing BD values. The

highest mean value of SOM and the lowest mean value of BD were obtained in soils over alluvium (Table 1) and further revealed that poorly drained alluvial soils are better accumulators of SOM. High SOM values are most likely to result in low BD values. Similar positive relationships were reported by Tisdall and Oades (1982) and Rawls et al. (2005) and contradict findings by Oliveira et al. (2018) that BD is unaffected by green manure. Others argue that organic matter does affect BD (Heuscher et al. 2005). An increase in organic matter oxidation rate is most likely to increase soil BD; for instance, poorly drained soils rich in accumulated organic matter have low BD compared to well-drained soils located in the upland. Conversely, an increase in green manure or SOM reduces BD (Boni et al. 1994; Parihar et al. 2016). However, this negative relationship was strongest in soils over limestone with higher  $r$  values ( $> 0.70$ ); especially those with Vertic properties as reported in Ofem et al.

**Table 1** Summary of descriptive statistics for the soils studied

Statistics	SOC g/kg	SOM	BD g/cm <sup>3</sup>	Ksat cm/h	Total P %	Air_P	CAP_P
Shale–limestone and sandstone intercalation (SLM) (IH1P1, IH1P2)							
Mean	9.86	16.94	1.57	34.99	51.41	8.97	42.41
Std	13.59	23.39	0.08	42.12	7.23	6.7	4.15
SE	5.14	8.84	0.031	15.92	2.73	2.53	1.57
Min	1.37	2.36	1.45	0.61	44.4	2.3	35
Max	40.3	69.32	1.66	106.28	65	21	47.8
CV	1.38	1.38	0.05	1.2	0.14	0.75	0.1
Alluvium (IH2P1, IH2P2, AI3P1, AI3P2, MF3P1, MF3P2)							
Mean	18.12	31.16	1.19	49.55	57.76	6.63	51.13
Std	23.24	39.98	0.32	85.44	10.01	5.67	10.49
SE	5.81	9.99	0.08	21.36	2.5	1.42	2.62
Min	1.03	1.77	0.53	0.49	45.1	2	38.1
Max	86.64	149.02	1.63	256.54	80.2	23.9	76.5
CV	1.28	1.28	0.27	1.72	0.17	0.86	0.21
Sandstone–limestone (SS) (AI1P1, AI1P2, MF1P2, MF2P2)							
Mean	7.01	12.06	1.43	43.67	48.8	4.71	45.07
Std	6.17	10.61	0.17	39.27	6.32	2.56	8.23
SE	1.54	2.65	0.043	9.82	1.58	0.64	2.06
Min	1.72	2.96	0.99	1.22	40.5	2.3	34.6
Max	21.96	37.77	1.66	126.67	63.4	11.3	60.7
CV	0.88	0.88	0.12	0.9	0.13	0.54	0.18
Limestone (LS) (AI2P1, AI2P2, MF1P1, MF2P1)							
Mean	7.72	13.28	1.35	16.67	49.56	5.3	44.13
Std	11.73	20.17	0.19	20.19	7.2	5.54	7.06
SE	3.13	5.39	0.051	5.4	1.92	1.48	1.89
Min	0.69	1.19	0.9	0.49	37.1	0.9	31.6
Max	46.34	79.7	1.6	75.52	62.2	23.5	59
CV	1.52	1.52	0.14	1.21	0.15	1.05	0.16

SOC soil organic carbon, SOM soil organic matter, BD bulk density, Ksat Saturated hydraulic conductivity, total P total porosity, Air\_P Air-filled porosity, Cap\_P capillary porosity, IH1, IH2, AI1, AI2, AI3, MF1, MF2, MF3 soil mapping units



**Fig. 1** Correlation between SOM and soil physical properties at  $p < 0.1$ , 1 and 5%. Bulk density (BD); saturated hydraulic conductivity (Ksat); total (Total\_P); air-filled (Air\_P); capillary porosities (Cap\_P)

(2020a), and alluvium ( $r = 0.578$ ), which had Loamic and Humic properties in the WRB system (Ofem et al. 2020a), and weakest in soils over SLM lithology.

### 3.1.2 Soil Organic Matter Versus Saturated Hydraulic Conductivity

Soil organic matter correlated moderately and positively with Ksat ( $r = 0.626$ ) ( $p < 0.001$ ) in the studied soils. Greater SOM results in higher Ksat because soil aggregate formation is linked to organic matter content (Beare et al. 1994). The presence of a considerable amount of organic matter ensures good aggregate and soil structural formation. This facilitates the movement of water through the soil. The highest value of SOM in soils over alluvium, which coincides with the highest value of Ksat, may further affirm their correlation. The soils over alluvium have either Aquic or Gleyic properties (Ofem et al. 2020a) expresses poorly drained soil conditions. Such conditions tend to encourage SOM deposition. Similar results have been reported, such that increased Ksat was

obtained through an increase in dairy manure application (Jiao et al. 2006; Eghball, 2002), and SOM in the Mediterranean region (Masri and Ryan 2006). However, the relationship is not always a straight positive correlation for any soil (Nemes et al. 2005). This indicates that SOM is most likely to increase if soil conditions that favor increased Ksat are created. Masri and Ryan (2006) recommended a legume rotation for improved Ksat. Generally, significant amounts of readily decomposed organic matter and enhanced nutrient release from such materials may improve physical soil conditions (Sanchez et al. 1989). A high positive correlation ( $r > 0.70$ ) was obtained between Ksat and SOM for soils over alluvium and LS, indicating greater certainty for the relationship than soils over SLM lithology. According to Saxton and Rawls (2006) and Yao et al. (2015), SOM is an important predictor of Ksat but strongly influenced by vegetations in the subtropics (Hao et al. 2019). For instance, irrespective of lithology, a higher mean value of 16.9 g/kg for SOM was obtained in the well-drained soils of the southern guinea savannah area compared to 12.06 and 13.28 g/kg obtained in the tropical

rainforest. This variation may be connected to the huge accumulation of litter in the oil palm and teak plantations.

### 3.1.3 Soil Organic Matter and Porosity

The correlation of SOM versus Total\_P resulted in a moderate positive correlation ( $r=0.44$ ) ( $p<0.01$ ) in the studied soils and implies increased SOM with an increase in Total\_P, but with moderate certainty in the positive relationship between SOM and Total\_P within sedimentary formations. Tisdall and Oades (1982). Nemes et al. (2005) obtained similar relationships as reported in this study. Boni et al. (1994), Whalen and Chang (2002), and Alves and Suzuki (2004) reported an increase in Total\_P by the use of green manure, dairy manure and successional cover crops. Similarly, Li et al. (2007) opinionated that a decrease in SOM will decrease porosity, reduced water and air storage. In soils over SS, the relationship was weak and positive. Soils over SS are high in the sand (Souza et al. 2019; Ofem et al. 2020b) and most likely to be well drained and more porous with a good supply of oxygen, and thus will most likely facilitate oxidation of organic matter (Bohn et al. 2001). This results in a high decomposition rate and low SOM accumulation.

Soil organic matter was very weakly correlated with Air\_P. This may suggest the indirect involvement of Air\_P in soil organic matter decay in the humid tropical region of southeast Nigeria. On the other hand, CAP\_P was positively moderately correlated ( $p<0.01$ ) with SOM ( $r=0.41$ ) in the studied soils. Total\_P and CAP\_P are highly correlated ( $r=0.80$ ), with each also positively correlated with SOM and both having the highest values in soils over alluvium. Soils over alluvium, therefore, exert a similar influence on SOM, Total\_P and CAP\_P. This implies that SOM increases with an increase in soil wetness conditions.

## 3.2 Principal Component Analysis

Principal Component Analysis (PCA) (Tables 2 and 3; Fig. 2) revealed that PC1 explained 54% of the variability in the dataset, while PC2 explained 22% of the variance between soils of diverse geological environments. PC1 was presented by the contribution of SOM, BD, Kstat, Total P, and CAP\_P, while PC2 was described by the contribution of Ksat, Air\_P and CAP\_P to their loadings (Table 3). The

**Table 2** Principal component contributions

Importance of components	PC1	PC2
Standard deviation	1.802	1.155
Proportion of variance	0.541	0.223
Cumulative proportion	0.541	0.764

**Table 3** Principal components correlation with variables

	PC1	PC2
SOM	<b>0.4091*</b>	0.1613
BD	<b>- 0.5119*</b>	0.1312
Ksat	0.3486	<b>0.4455*</b>
Total_P	<b>0.4872*</b>	0.0134
Air_P	0.0485	<b>0.7559*</b>
CAP_P	<b>0.4569*</b>	<b>- 0.4321*</b>

NB: Bolded values showed the variables contributing more to each PC

\*Contribution to each PC

points outside the ellipses are outliers of each of the geological environments.

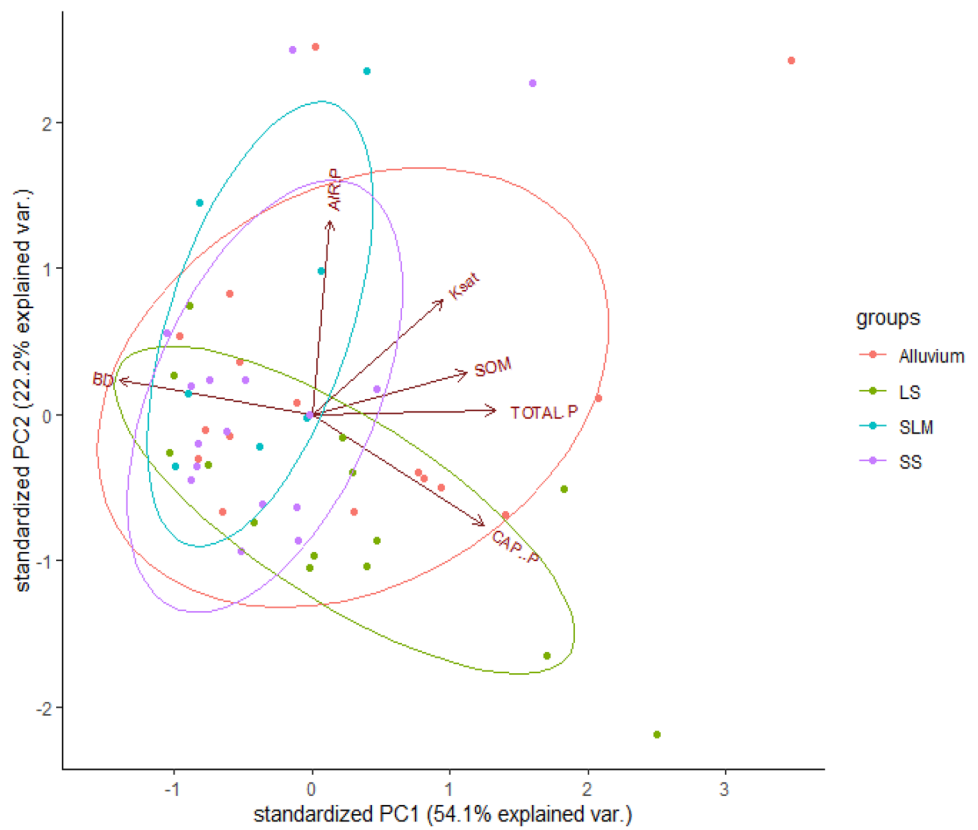
All the soil properties were significantly influenced by SOM ( $p<0.01$ , 0.001) under alluvial deposits except Air\_P. Similarly, BD ( $r=-0.783$ ), Ksat, Total\_P and Cap\_P ( $r>0.54$ ) were affected by SOM under LS. SOM was reportedly positively inter-related with Ksat and inversely with BD in soils formed over SS, while SOM had no influence on the properties for soils over SLM. The PCA result reconfirmed the correlation matrix output (Fig. 1).

## 3.3 SOM Prediction

Presented in Table 4 is the result of the four stepwise linear regression models for SOM prediction. Leapforward yield (RMSE = 12.51%,  $R^2=0.53$ , MAE = 8.68%), Leapbackward gave (RMSE = 11.50%,  $R^2=0.58$ , MAE = 8.48%), leapseq yielded (RMSE = 12.51%,  $R^2=0.53$ , MAE = 8.68%) and lmStepAIC function produced (RMSE = 13.24%,  $R^2=0.54$ , MAE = 9.56%). The results revealed that the best performing function for SOM prediction is the leapbackward function since it produced the lowest error with a high coefficient of determination value. However, all the model functions were within the acceptable prediction range ( $R^2 \geq 0.50-0.75$ ) as proposed by Li et al. (2016). These results suggest that prediction of SOM may vary depending on the method/functions adopted. The backward elimination (leapbackward) likewise, the rest functions procedure identified the best model as having BD\*\* and Ksat\*\*\*, respectively.

According to Sakin (2012), BD is closely related to SOC by storing large amounts of SOM. Compacted soil may contain more SOM, as it will occupy less space and more SOM per volume of soil, and the SOM in compacted soil is essentially "locked away". In contrast, soils that are not compacted have more contact with the air in the soil pores and so can be mineralized more efficiently and used as plant nutrients or leached. This relationship has been reported to aid the estimation of BD from SOM and vice versa (Perie and Ouimet, 2008). The study by Adams (1973) revealed that SOM had

**Fig. 2** Principal component analysis of the variables grouped by geological environment. Bulk density (BD); saturated hydraulic conductivity (Ksat); total (Total\_P); air-filled (Air\_P); capillary porosities (Cap\_P)



**Table 4** Prediction of soil organic matter (SOM) via various stepwise linear functions

LM Functions	RMSE (%)	R <sup>2</sup>	MAE (%)	Equations	Variable of importance
LeapForward	12.51	0.53	8.68	SOM = 38.2 - 23.6BD + 0.13Ksat	BD**, Ksat***
LeapBackward	<b>11.50</b>	<b>0.58</b>	<b>8.48</b>	<b>SOM = 38.2 - 23.6BD + 0.13Ksat</b>	<b>BD**, Ksat***</b>
LeapSeq	12.51	0.53	8.68	SOM = 38.2 - 23.6BD + 0.13Ksat	BD**, Ksat***
LmStepAIC	13.24	0.54	9.56	SOM = 38.2 - 23.6BD + 0.13Ksat	BD**, Ksat***

*p* = 0.001 '\*\*\*'; 0.01 '\*\*', Bold gave a good model fit

a dominant effect on both bulk and actual densities of soil in podzolic soil's organic and eluvial horizons.

Similarly, as shown in the correlation matrix, Ksat gave a higher correlation with SOM than BD; this was captured in all the four linear regression functions used in this study. The regression result confirms that an increase in SOM in the soil will result in a proportional increase in Ksat. This is because Ksat describes the capability of the bulk soil to transmit water when subjected to a hydraulic gradient. This is expressed by the volume of water flowing per unit area of bulk soil per unit time (Kosugi et al. 2002). Also, the result in this study is similar to the report of Ankenbauer and Loheide (2017). They reported an  $R^2 = 0.625$  in predicting SOM via volumetric water content at saturation in the meadow of the Sierra Nevada.

Generally, organic matter has been reported to significantly influence soil water retention and BD (Rawls et al.

2003; Olness and Archer, 2005; Saxton and Rawls, 2006). In contrast, other studies have reported that SOM is not necessary to estimate soil water retention properties accurately (Zhuang et al. 2001). However, in a dissimilar geological environment like this study, where the soils are predominantly similar in texture (Ofem et al. 2020a) and SOM content from 12.06 to 31.6 g/kg, SOM can easily be estimated via BD and Ksat. This is because SOM exerts a substantial control on surface water retention and BD variability.

### 4 Conclusions

Soil organic matter is most likely to increase when favorable conditions for increased Ksat and porosity except Air\_P, which did not influence SOM. Irrespective of geological material, BD decreases when SOM increases. The Ksat of

soils over limestone (LS) and alluvium and BD of soils over LS had the strongest relationships with SOM with  $r > 0.70$ . However, air-filled (Air\_P) porosity had no significant association with SOM and is most likely to have little effect on its decomposition in sedimentary geological environments. Farmers must put in place measures to regulate soil moisture (mulching and drainage), particularly in the sedimentary geological environment, which affects SOM. PC1 and PC2 contributed 74.38% of the total variance in the dataset of soils over diverse geological environments. The grouping pattern in the PCA explained that alluvial deposits influence most soil characteristics in this present study.

All the selected stepwise linear regression functions in the R environment performed the same as they fell within acceptable prediction criteria ( $R^2 = 0.50\text{--}0.75$ ). However, the best performing model function was *leapbackward*, which produced a smaller error when compared with others. The models selected BD and Ksat as the most important variables to explain the SOM variability in diverse sedimentary geology. The reason behind this result could not be presented at the time of this study; however, it could be interesting to access these functions with more variables and large sample densities. Therefore, we propose an increase in sample density per lithological make-up and the incorporation of soil properties known in works of literature to be influenced by SOM. This is to verify the performance of the *leapbackward* function over other functions, including the conventional *lmStepAIC* algorithm.

**Data availability** Not applicable.

## Declarations

**Conflict of interest** The authors declare no conflict of interest regarding this work.

**Ethical approval** Not applicable.

**Consent to participate** All authors gave their consent.

**Concent for publication** All authors gave their approval.

## References

- Adams WA (1973) The effect of organic matter on the bulk and true densities of some uncultivated podzolic soils. *J Soil Sci* 24(1):10–17
- Ali S, Begum F, Hayat R, Bohannam BJM (2017) Variation in soil organic carbon stock in different land uses and altitudes in Bagrot Valley, Northern Karakoram. *Acta Agric Scand B Soil Plant Sci* 67:551–561
- Alves MC, Suzuki LE (2004) Influência de diferentes sistemas de manejo do solo na recuperação de suas propriedades físicas. *Acta Sci Agron* 26:27–34
- Ankenbauer KJ, Loheide SP (2017) The effects of soil organic matter on soil water retention and plant water use in a meadow of the Sierra Nevada, CA. *Hydrol Process* 31(4):891–901
- Beare MH, Hendrix PF, Coleman DC (1994) Water-stable aggregates and organic matter fractions in conventional and no-tillage soils. *Soil Sci Soc Am J* 58:777–786
- Belkhir L, Narany TS (2015) Using multivariate statistical analysis, geostatistical techniques, and structural equation modeling to identify spatial variability of groundwater quality. *Water Resour Manag* 29:2073–2089
- Blanco-Canqui H, Shapiro CA, Wortmann CS, Drijber RA, Mamo M, Shaver TM, Ferguson RB (2013) Soil organic carbon: The value to soil properties. *J Soil Water Conserv* 68(5):129A–134A. <https://doi.org/10.2489/jswc.68.5.129A>
- Bohn H, McNeal B, O'Connor G (2001) *Soil Chemistry*, 3rd edn. Wiley Interscience, New York
- Boni, N. R., Espindola, C. R, Guimarães, E. C. (1994). Uso de leguminosas na recuperação de um solo decapitado. In: Simpósio nacional sobre recuperação de áreas Degradadas, Curitiba. pp. 563–568. (Fundação de Pesquisas Florestais do Paraná)
- Bonini, C. S. B. and Alves, M. C. (2010). Relation between soil organic matter and physical properties of a degraded Oxisol in recovery with green manure, lime and pasture. 19th World Congress of Soil Science, Soil Solutions for a Changing World 1—6 August 2010, Brisbane, Australia.
- Cardinael R, Chevallier T, Cambou A, Beral C, Barthès BG, Dupraz C, Durand C, Kouakoua E, Chenu C (2017) Increased soil organic carbon stocks under agroforestry: A survey of six different sites in France. *Agric Ecosyst Environ* 2017(236):243–255. <https://doi.org/10.1016/j.agee.2016.12.011>
- Eghball B (2002) Soil properties as influenced by phosphorus and nitrogen based manure and compost applications. *Agron J* 94:128–135
- Ekwueme BN (1987) Structural orientations and Precambrian deformational episodes of Uwet Area, Oban Massif, SE Nigeria. *Precamb Res* 34:269–289
- Fatoye FB, Gideon YB (2013) Geology and occurrences of limestone and marble in Nigeria. *J Nat Sci Res* 3(11):60–65
- Ghimire P, Kafle G, Bhatta B (2018) Carbon stocks in *Shorea robusta* and *Pinus roxburghii* forests in Makawanpur district of Nepal. *J AFU* 2:241–248
- Hao M, Zhang J, Meng M, Chen HYH, Guo X, Liu S, Ye L (2019) Impacts of changes in vegetation on saturated hydraulic conductivity of soil in subtropical forests. *Sci Rep* 9:8372. <https://doi.org/10.1038/s41598-019-44921-w>
- Hastie T, Tibshirani R, Friedman J (2008) *The elements of statistical learning: data mining, inference, and prediction*. Springer Science and Business Media, Berlin
- Heuscher SA, Brandt CC, Jardine PM (2005) Using soil physical and chemical properties to estimate bulk density. *Soil Sci Soc Am J* 69:51–56. <https://doi.org/10.2136/sssaj2005.0051a>
- James G, Witten D, Hastie T, Tibshirani R (2014) *An introduction to statistical learning: with applications in R*. Springer Publishing Company Incorporated, Berlin
- Jiao Y, Whalen JK, Hendershot WH (2006) No-tillage and manure applications increase aggregation and improve nutrient retention in a sandy-loam soil. *Geoderma* 134:24–33
- Jiao S, Li J, Li Y, Xu Z, Kong B, Li Y, Shen Y (2020) Variation of soil organic carbon and physical properties in relation to land uses in the Yellow River Delta. *China Scientific Reports* 10:20317. <https://doi.org/10.1038/s41598-020-77303-8>
- John K, Abraham Isong I, Michael Kebonye N, Okon Ayito E, Chapman Agyeman P, Marcus Afu S (2020) Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* 9(12):487



- Kafle G (2019) Vertical distribution of soil organic carbon and nitrogen in a tropical community forest of Nepal. *Int J Forest Res.* <https://doi.org/10.1155/2019/3087570>
- Kebonye NM, John K, Chakraborty S, Agyeman PC, Ahado SK, Eze PN, Borůvka L (2020) Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma* 384:114792
- Kogel-Knabner I (2002) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biol Biochem* 34(2002):139–162
- Komarov A, Chertov O, Bykhovetsa S, Shaw C, Nadporozhskaya M, Frolova P, Shashkova M, Shanina V, Grabarnika P, Priputinina I (2017) Romul-Hum model of soil organic matter formation coupled with soilbiota activity. I. Problem formulation, model description, and testing. *Ecol Model* 345:113–124
- Kosugi KI, Hopmans JW, Dane JH (2002) Parametric models. *Methods Soil Analysis* 5:739–757
- Lei Z, Yu D, Zhou F, Zhang Y, Yu D, Zhou Y, Han Y (2019) changes in soil organic carbon and its influencing factors in the growth of *Pinus sylvestris* var. *mongolica* plantation in Horqin Sandy Land, northeast china. *Sci Rep* 9:16453. <https://doi.org/10.1038/s41598-019-52945-5>
- Li XG, Li FM, Zed R, Zhan ZY, Bhupinderpal S (2007) Soil physical properties and their relations to organic carbon pools as affected by land use in an alpine pastureland. *Geoderma* 139:98–105
- Li L, Lu J, Wang S, Ma Y, Wei Q, Li X, Cong R, Ren T (2016) Methods for estimating leaf nitrogen concentration of winter oilseed rape (*Brassica napus* L.) using in situ leaf spectroscopy. *Ind Crops Prod* 91:194–204
- Li Z, Liu C, Dong Y, Chang X, Nie X, Liu L, Zeng G (2017) Response of soil organic carbon and nitrogen stocks to soil erosion and land use types in the Loess hilly–gully region of China. *Soil Tillage Res* 166:1–9
- Liu S, An N, Yang J, Dong S, Wang C, Yin Y (2015) Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China. *CATENA* 133:137–144. <https://doi.org/10.1016/j.catena.2015.05.0100>
- Masri Z, Ryan J (2006) Soil organic matter and related physical properties in a mediterranean wheat-based rotation trial. *Soil Tillage Res* 87:146–154. <https://doi.org/10.1016/j.still.2005.03.003>
- Muñoz-Rojas M, Abd-Elmabod SK, Zavala LM, DelaRosa D, Jordán A (2017) Climate change impacts on soil organic carbon stocks of Mediterranean agricultural areas: a case study in Northern Egypt. *Agr Ecosyst Environ* 238:142–152. <https://doi.org/10.1016/j.agee.2016.09.001>
- Nemes A, Rawls WJ, Pachepsky YA (2005) Influence of organic matter on the estimation of saturated hydraulic conductivity. *Soil Sci Soc Am J* 69:1330–1337. <https://doi.org/10.2136/sssaj2004.0055>
- Obi ME (2000) Laboratory manual for soil physics. Department of Soil Science, University of Nigeria, Nsukka, p 34
- Ofem KI, Abua SO, Umeugokwe CP, Ezeaku VI, Akpan-Idiok AU (2020a) Characterization and suitability evaluation of soils over sandstone for cashew (*Anacardium occidentale* L.) production in a Nigerian Southern Guinea Savanna. *IJSRP* 10(7):353–368
- Ofem KI, Asadu CLA, Ezeaku PI, Kingsley J, Eyong MO, Katerina V, Václav T, Karel N, Ondrej D, Vít P (2020b) Genesis and classification of soils over limestone formations in a Tropical Humid Region. *Asian J Sci Res* 13:228–243
- Oliveira FCC, Ferreira GWD, Souza JLS, Vieira MEO, Pedrotti A (2018) Soil physical properties and soil organic carbon content in northeast Brazil: long-term tillage systems effects. *Soils Plant Nutr.* <https://doi.org/10.1590/1678-992X-2018-0166>
- Olness A, Archer D (2005) Effect of organic carbon on available water in soil. *Soil Sci* 170(2):90–101
- Parihar CM, Yadav MR, Jat SL, Singh AK, Kumar B, Pradhan S, Chakraborty D, Jat ML, Jat RK, Saharawat YS, Yadav OP (2016) Long term effect of conservation agriculture in maize rotations on total organic carbon, physical and biological properties of a sandy loam soil in north-western Indo-Gangetic Plains. *Soil Tillage Res* 161:116–128. <https://doi.org/10.1016/j.still.2016.04.001>
- Perie C, Ouimet R (2008) Organic carbon, organic matter and bulk density relationships in boreal forest soils. *Can J Soil Sci* 88(3):315–325
- R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria [online]. <https://www.r-project.org/>. Accessed 13 May 2020.
- Rawls WJ, Pachepsky YA, Ritchie JC, Sobecki TM, Bloodworth H (2003) Effect of soil organic carbon on soil water retention. *Geoderma* 116(1–2):61–76
- Rawls WJ, Nemes A, Pachepsky YA (2005) Effect of soil organic matter on soil hydraulic properties. In: Pachepsky YaA, Rawls WJ (eds) Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, pp 95–114
- Saint-Laurent D, Gervais-Beaulac V, Paradis R, Arsenault-Boucher L, Demers S (2017) Distribution of soil organic carbon in riparian forest soils affected by frequent floods (Southern Québec, Canada). *Forests* 8:124. <https://doi.org/10.3390/f8040124>
- Sakin E (2012) Organic carbon organic matter and bulk density relationships in arid-semi arid soils in Southeast Anatolia region. *Afr J Biotech* 11(6):1373–1377
- Sambo, E. E., Ufoegbune, G. C., Eruola, A. O. & Ojekunle, O. Z. (2016). Impact of Rainfall Variability on Flooding of Rivers in Cross River Basin, Nigeria. Nigerian Meteorological Society (NMETS), "Climate Variability And Change: Impact, Science, Innovation And Policy" at Federal College of Education Osiele, Abeokuta. 21—24 November, 2016
- Sanchez PA, Palm CA, Szott LT, Cuevas E, Lal R (1989) Organic input management in tropical agrosystems. In: Coleman DC, Oades JM, Uehara G (eds) Dynamics of soil organic matter in tropical ecosystems. Honolulu, Capi' tulo 2, pp 125–152 (**ISBN 0-8248-1251-4**)
- Saxton KE, Rawls WJ (2006) Soil water characteristics estimates by texture and organic matter for hydrologic solutions. *Soil Sci Soc Am J* 70:1569–1578. <https://doi.org/10.2136/sssaj2005.0117>
- Silva SHG, Teixeira AFS, Menezes MD, Guilherme LRG, Moreira FMS, Curi N (2017) Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). *Ciênc Agrotec* 41(6):648–664. <https://doi.org/10.1590/1413-70542017416010317>
- Soil Survey Staff (SSS) (2014). Kellogg soil survey laboratory methods manual. Soil Survey Investigations Report No. 42, Version 5.0. R. Burt and Soil Survey Staff (ed.). US Department of Agriculture, Natural Resources Conservation Service. P. 1001.
- Souza DF, Barbosa RS, da Bezerra Silva YJ, de Soares Moura MC, de Porto Oliveira R, Martins V (2019) Genesis of sandstone-derived soils in the Cerrado of the Piauí State Brazil. *Rev Ambient Água.* <https://doi.org/10.4136/1980-993X> (**ISSN 1980-993X**)
- Takata Y, Funakawa S, Akshalov K, Ishida N, Kosaki T (2007) Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. *Soil Sci Plant Nutr* 53(3):289–299. <https://doi.org/10.1111/j.1747-0765.2007.00142.x>
- Tisdall JM, Oades JM (1982) Organic matter and water-stable aggregates in soils. *J Soil Sci* 33:141–163
- Whalen JK, Chang C (2002) Macroaggregate characteristics in cultivated soils after 25 annual manure applications. *Soil Sci Soc Am J* 66:1637–1647
- Yao R, Yang J, Wu D, Li F, Gao P, Wang X (2015) Evaluation of pedotransfer functions for estimating saturated hydraulic conductivity in coastal salt-affected mud farmland. *J Soils Sediments* 15:902–916. <https://doi.org/10.1007/s11368-014-1055-5>

- Zhao W, Hu Z, Yang H, Zhang L, Guo Q, Wu Z, Liu D, Li S (2016) Carbon density characteristics of sparse *Ulmus pumila* forest and *Populus simonii* plantation in Onqin Daga Sandy Land and their relationships with stand age. *Chin J Plant Ecol* 40:318–326
- Zhou W, Han G, Liu M, Zeng J, Liang B, Liu J, Qu R (2020) Determining the distribution and interaction of soil organic carbon, nitrogen, pH and texture in soil profiles: a case study in the langcangjiang River Basin, Southwest China. *Forests* 11:532. <https://doi.org/10.3390/f11050532>
- Zhuang J, Jin Y, Miyazaki T (2001) Estimating water retention characteristic from soil particle-size distribution using a non-similar media concept. *Soil Sci* 166(5):308–321



## Assessing the impact of sampling strategy in random forest-based predicting of soil nutrients: a study case from northern Morocco

Kingsley John, Yassine Bouslihim, Abdelkrim Bouasria, Rachid Razouk, Lahcen Hssaini, Isong Abraham Isong, Samir Ait M'barek, Esther O. Ayito & Gare Ambrose-Igho

To cite this article: Kingsley John, Yassine Bouslihim, Abdelkrim Bouasria, Rachid Razouk, Lahcen Hssaini, Isong Abraham Isong, Samir Ait M'barek, Esther O. Ayito & Gare Ambrose-Igho (2022): Assessing the impact of sampling strategy in random forest-based predicting of soil nutrients: a study case from northern Morocco, Geocarto International, DOI: [10.1080/10106049.2022.2048091](https://doi.org/10.1080/10106049.2022.2048091)

To link to this article: <https://doi.org/10.1080/10106049.2022.2048091>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 16 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 721







View related articles [↗](#)



View Crossmark data [↗](#)

# Assessing the impact of sampling strategy in random forest-based predicting of soil nutrients: a study case from northern Morocco

Kingsley John<sup>a</sup>, Yassine Bouslihim<sup>b#</sup> , Abdelkrim Bouasria<sup>c</sup> , Rachid Razouk<sup>b</sup> , Lahcen Hssaini<sup>b</sup> , Isong Abraham Isong<sup>d</sup>, Samir Ait M'barek<sup>e</sup>, Esther O. Ayito<sup>d</sup> and Gare Ambrose-Igho<sup>f</sup>

<sup>a</sup>Faculty of Agrobiological, Food, and Natural Resources, Department of Soil Science and Soil Protection, Czech University of Life Sciences, Prague, Czech Republic; <sup>b</sup>National Institute for Agricultural Research (INRA), Rabat, Morocco; <sup>c</sup>Faculty of Sciences, Department of Geology, Chouaib Doukkali University, El Jadida, Morocco; <sup>d</sup>Department of Soil Science, Faculty of Agriculture, Forestry and Wildlife Resources Management, University of Calabar, Calabar, Nigeria; <sup>e</sup>Faculty of Sciences and Techniques Settat, Hassan First University, Settat, Morocco <sup>f</sup>Department of Geoscience, University of Nevada, Las Vegas, NV, USA

## ABSTRACT

In this work, we tested different combinations of sampling strategies, random sampling and conditioned Latin Hypercube sampling (cLHS) and sample ratios (10% = 147 and 25% = 368) to predict soil phosphorus and potassium contents, previously estimated using standard laboratory protocols. Other environmental covariates, used as input data for prediction, were obtained from different sources (multispectral Landsat-OLI 8 image, WorldClim database, ISRIC soil database, and ASTER-GDEM). Our findings showed that random sampling was suitable for predicting phosphorus, while the conditioned Latin Hypercube sampling was suitable for predicting potassium. Furthermore, we observed that when the sample ratio increased from 10 to 25%, model accuracy improved in random sampling and cLHS for phosphorus and potassium prediction. However, before generalizing these findings, we recommend that further studies be conducted under different conditions (climate, soil types and parent materials) and testing other sample ratios to determine the best sampling strategy with the optimum ratio to predict soil nutrients better.

## ARTICLE HISTORY

Received 7 November 2021  
Accepted 21 February 2022

## KEYWORDS

Predictive soil mapping; sampling strategies; soil nutrients; soil management; machine learning; Mediterranean region

## 1. Introduction

Phosphorus and potassium are essential macronutrients for plant growth and nitrogen (Morgan and Connolly 2013). Phosphorus is released into the soil solution by decomposing rock fragments and mineralizing organic matter (Gumiere et al. 2019). Potassium is

**CONTACT** Kingsley John  [johnk@af.czu.cz](mailto:johnk@af.czu.cz)

<sup>#</sup>Mohammed VI Polytechnic University (UM6P), International Water Research Institute (IWRI), Ben Guerir, Morocco. This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

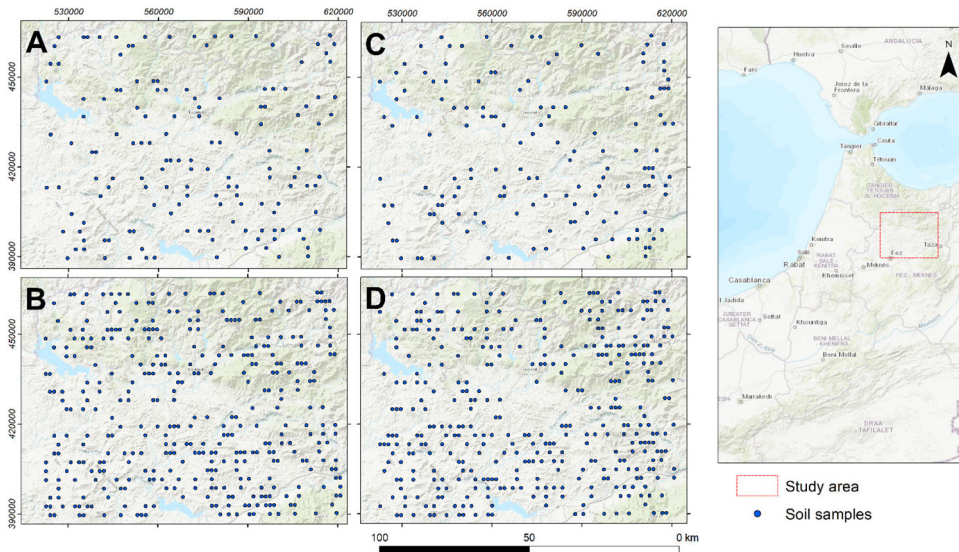
made available in the soil solution through exchangeable K (Öborn et al. 2005). Thus, plants regularly take them up from the soil solution. Soil, phosphorus, and potassium regulate plant cell processes such as energy, photosynthesis, and sugar transformation (Ramaekers et al. 2010; Kadarwati 2020). Potassium plays a vital role in yield and quality improvement (Marschner and Dell 1994; Oosterhuis et al. 2014), while phosphorus plays a significant role in root development, nutrient uptake and crops growth (Abbasi et al. 2008). They are both required in large amounts for any crop development. Therefore, any soil and land management leading to their low availability may result in unhealthy crops, low yield, or even the death of crops.

Soil nutrient maps provide the spatial distribution of essential nutrients for precision agriculture and identify areas for intervention. Furthermore, the accuracy and quality of these essential soil nutrient maps (such as phosphorus and potassium) generally depend on the sampling point distribution, size and the model adopted for mapping. For this purpose and to develop reliable predictions of a targeted soil nutrient efficiently and cost-effectively, the minimum sample size and the most appropriate sampling strategy must first be evaluated.

Different sampling strategies have been proposed in predicting various targeted soil properties. For example, Wollenhaupt et al. (1994) suggested two-dimensional grid sampling strategies for mapping phosphorus and potassium for site-specific recommendations. However, this sampling strategy is limited since it requires large observation points to make accurate soil nutrient recommendations, especially in highly complex sites. Unfortunately, the extensive labor needed in soil sample collection and laboratory analysis costs makes this grid sampling strategy unfeasible on a large scale (Varvel et al. 1999; Kozar et al. 2002; Higo et al. 2015). Other sampling strategies used in soil nutrients prediction include classical random sampling, stratified random sampling (Brus and De Gruijter 1997), and conditioned Latin Hypercube Sampling (cLHS) (Brungard and Boettinger 2010).

In a simple random sampling strategy, a pre-determined number of sample positions are randomly selected from the area, and the selection probabilities are equal and independent of each other. This is done by obtaining the geographic coordinates of each sample location from a random number generator or random number table. In stratified random sampling, the area is first divided into several sub-areas, called layers, and then simple random sampling is used for each layer (Brus and De Gruijter 1997). Furthermore, Brus and De Gruijter (1997) outlined that the choice of sampling depends on many factors, proposing more comparative research on sampling strategies. On the other hand, the conditioned Latin conditioned Hypercube sampling is a stratified random sampling, which accurately represents the variability of environmental covariates in the feature space (Brungard and Boettinger 2010). This sampling strategy has been used extensively in digital soil mapping (DSM) projects around the world, as recently reported in the last five years by Sun et al. (2017) in China, Jeong et al. (2017) in South Korea, Scarpone et al. (2016) in Canada and Thomas et al. (2015) in Australia. In addition, Minasny and McBratney (2006) outlined that cLHS was most effective in replicating the distribution of the variables.

Sampling design and sample size are the most vital criteria to consider in predicting soil parameters that vary spatially based on the vast heterogeneity of the soil environments. Also, prediction accuracy is partly influenced by the sample size and spatial positions of the sampling points with weights of the target property used to build the model. Besides that, little has been investigated on optimal sampling strategies for mapping using the random forest to the best of our knowledge. Therefore, this paper compared two different sampling strategies, random sampling and conditioned Latin hypercube sampling, each with two different sample sizes to achieve the optimum prediction of soil nutrients



**Figure 1.** Map of the study area showing the location of the sampling points, using conditioned Latin sampling (cLHS) with A: cLHS\_10 ( $n = 147$ ), B: cLHS\_25 ( $n = 368$ ), C: rand\_10 ( $n = 147$ ) and D: rand\_25 ( $n = 368$ ).

based on a random forest model for proper soil fertilizer and application of other soil input materials in the Mediterranean region of Morocco.

## 2. Materials and methods

### 2.1. Site description

The study area is located in the province of Taounate in northern Morocco ( $34^{\circ}47'N$ ,  $4^{\circ}4.4'W$  and  $34^{\circ}05'N$ ,  $5^{\circ}10.3'W$ ), displayed as 7700 square kilometers ( $100 \times 77$  km) (Figure 1). It mainly belongs to the mountainous region of the central Rif. In addition, it includes a part of the Atlas Mountains in the northwest. Generally, the area covered by this study varies from 78 to 1969 meters. Its geological formations consist of a Jurassic-Cretaceous series of marl overcome molassic formations composed of sandstone and conglomerates (Mesrar et al. 2017). According to World Reference Base for Soil Resources (FAO Classification System), the region's soils are moderately weathered and classified within Xeralsfs and Luvisols. They consist of very deep, somewhat poorly drained soils formed in sandy outwash, glaciolacustrine, or eolian deposits on outwash plains, lake plains, and dunes (FAO/ISRIC/ISSS 2006).

The region's climate is of Mediterranean type, with rainy winter and dry summer. Rainfall is irregular throughout the year, with an annual average of about 650 mm, mostly occurring from October to February. The average yearly temperature is  $17^{\circ}C$ . The average maximum temperature of the hottest month (July) is about  $34.2^{\circ}C$ , while the average minimum temperature of the coldest month (January) is  $0.5^{\circ}C$  (Allali et al. 2020; Rezouki et al. 2020).

The area is characterized by diversified vegetation cover, mainly including cereals (wheat and barley), fruit trees (olive, almond, fig), legumes (beans, chickpeas, lentils), and fodder crops. Recently, the area's development has encountered problems with cannabis cultivation, of which 3,000 hectares have been razed, and a massive plan to plant drought-tolerant trees, such as olive and carob, have been launched (Kradi 2012).

## 2.2. Sampling strategies, sample size and laboratory protocol

In total, 1470 points were collected to a 0–40 cm depth over two months (November and December 2013), using a 1 km grid to cover the entire study area. During the sampling campaigns, all sampling and transport standards were respected. Available K and P were analyzed by extraction method using ammonium acetate (1:10) and molybdate ammonium (1:20), respectively.

The 1470 samples were used to extract a set of parameters (topography, remote sensing indices, climate, soil, section 2.3) for each point. This final database, containing all the sampling points and all environmental covariates values, adopted two different sampling strategies with different sizes (10% = 147 samples and 25% = 368 samples). The first is the random sampling method, and the second is conditioned Latin hypercube sampling (cLHS). The first set of sample sizes with sampling strategies is represented as *clhs\_10* and *rand\_10* ( $n=147$  samples). At the same time, the set of the second set of sample sizes and sampling strategies are defined as *clhs\_25* and *rand\_25* ( $n=368$  samples) (Figure 1).

Both methods were conducted in python using Google Collab and other libraries such as *pandas*, *numpy* and *clhs*. We used 10,000 iterations to perform the functions of both sampling strategies, as recommended by Malone et al. (2019). Furthermore, for the cLHS, which requires auxiliary data for execution, the samples were subset using the following auxiliary covariates: slope, elevation, Normalized Difference Vegetation Index (NDVI), and Soil Adjusted Vegetation Index (SAVI).

## 2.3. Auxiliary covariates sources and preparation

In this study, the covariates used can influence the spatial distribution of phosphorus and potassium content in soils. We obtained soil properties such as clay, silt, and sand, cation exchange capacity (CEC) and bulk density (BD) from the ISRIC soil database. In addition, four bioclimatic parameters representing temperature variation [annual mean temperature (*bio\_1*), max temperature of warmest month (*bio\_5*) and min temperature of coldest month (*bio\_6*)] and precipitation [annual precipitation (*bio\_12*)] were obtained from the WorldClim database version 2 (Fick and Hijmans 2017). These data are available in GeoTiff (.tif) format with a resolution of 10 minutes ( $\sim 340$  km<sup>2</sup>). Also, from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)-Global Digital Elevation Model (GDEM) with a resolution of 30 m, terrain attributes such as elevation, slope, profile curvature, plan curvature, Multi-resolution Valley Bottom Flatness (MrVBF), Multi-resolution Ridge Top Flatness (MrRTF), Topographic Wetness Index (TWI), convergence index and aspect were extracted. Vegetation indices have shown their importance in predicting different soil parameters (Mirzaee et al. 2016; Bouslihimi et al. 2021). For this reason, the Landsat-8 OLI/TIRS image with a spatial resolution of 30 m was used to extract the six parameters based on the equations presented in Table 1. The calculated indices are as follows: Normalized Difference Vegetation Index (NDVI), Transformed Normalized Difference Vegetation Index (TNDVI), Soil Adjusted Vegetation Index (SAVI), Ratio Vegetation Index (RVI), Difference Vegetation Index (DVI), and Chlorophyll Vegetation Index (CVI). Since the study area is within two scenes' limits (path = 200/201 and row = 36), two Landsat 8 images were used to calculate all required parameters. Furthermore, to estimate the various parameters, the pre-processing of all used bands, treatment, and analysis were done in the ArcGIS program. All the covariates

**Table 1.** Auxiliary covariates applied in the modelling regime.

Covariates	Sources	Resolution	Resample resolution
<i>Soil properties</i>			
Particle size fractions (clay, silt, sand)	ISRIC soil database	250 m	30 m
Cation exchange capacity (CEC)			
Bulk density (BD)			
<i>Bioclimatic parameters</i>			
Annual mean temperature (bio_1)	WorldClim database version 2	10 minutes (~ 340 km <sup>2</sup> )	30 m
Max temperature of the warmest month (bio_5)			
Min temperature of the coldest month (bio_6)			
Annual precipitation (bio_12)			
<i>Terrain attributes</i>			
Elevation; slope; profile curvature; plan curvature, Multi-resolution Valley Bottom Flatness (MrVBF); Multi-resolution Ridge Top Flatness (MrRTF); Topographic Wetness Index (Topographi); convergence index and aspect	ASTER-GDEM	30 m	30 m
<i>Remote sensing indices</i>			
Normalized Difference Vegetation Index (NDVI)	Landsat OLI 8	30 m	m
Transformed Normalized Difference Vegetation Index (TNDVI)			
Soil Adjusted Vegetation Index (SAVI)			
Ratio Vegetation Index (RVI) Difference Vegetation Index (DVI)			
Chlorophyll Vegetation Index (CVI)			
<i>Coordinates</i>		30 m	30 m
x			
y			

with coarser resolution were downscaled to 30 m pixel size using the nearest neighbor function in the ArcGIS program.

## 2.4. Experimental design

Figure 2 gives a general idea of the methodology adopted in the present study, starting from the preparation of the independent variables, passing through the sampling strategies for the sample points selection; the method adopted for the implementation, the uncertainty evaluation and the model performance and finishing with the nutrient mapping.

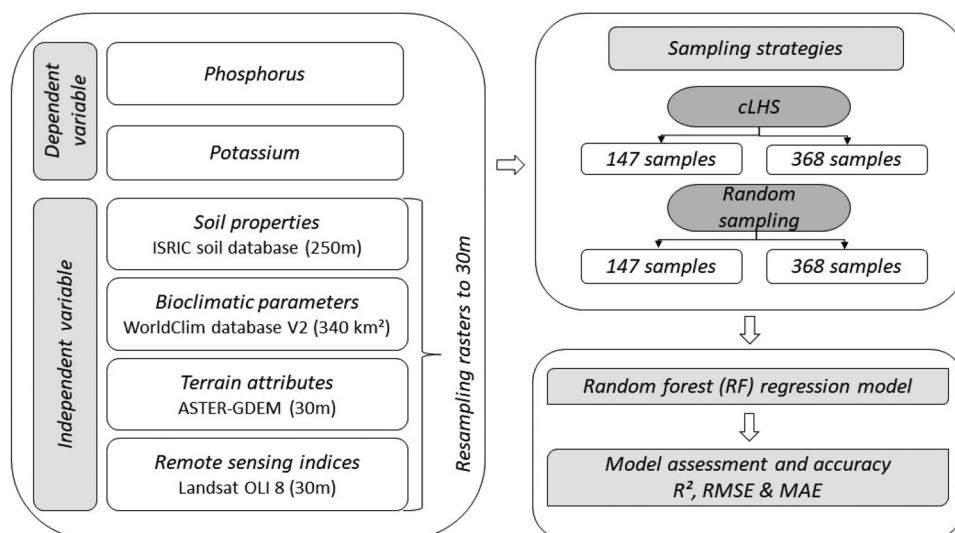
### 2.4.1. Random forest (RF) regression model

RF is a classification and regression tree (CART)-based machine learning technique (Breiman 2001). It uses the bootstrap resampling method to extract multiple samples from the original samples. It then models each bootstrap sample decision tree and combines multiple decision trees for classification and regression during the training process. The turning parameters was implemented under  $mtry = p/3$  (i.e.,  $30/3 = 9$ ), where  $p$  is the total number of variables and  $ntree = 1000$ . The final prediction is the average of all tree prediction results. This was performed in an R environment.

### 2.4. Dataset partitioning and model assessment

As earlier stated in section 2.2, we obtained two sets of samples,  $N = 147$  and  $368$ , via different sampling regimes. Therefore, for modelling, we randomly partitioned the data into two separate datasets (i.e., a calibration dataset and a validation dataset), with a ratio of





**Figure 2.** Research methodology flowchart.

70% to 30%. As a result, the calibration dataset for cLHS<sub>10</sub> and rand<sub>10</sub> was composed of 104 and 43 soil samples for calibration and validation, respectively. While for cLHS<sub>25</sub> and rand<sub>25</sub>, the calibration dataset was composed of 260 soil samples, and the validation set 108 soil samples. This was performed in an R environment with the function *createDataPartition* in the caret package (Hyndman and Athanasopoulos 2018).

Model accuracy and performance were evaluated using root mean square error (RMSE), coefficient of determination ( $R^2$ ), and mean absolute error (MAE).

### 3. Results and discussion

#### 3.1. Summary statistics

The summary statistics of sampling strategy and sample size in predicting soil phosphorus ( $P_2O_5$ ) and potassium ( $K_2O$ ) are reported in Tables 2 and 3. With an emphasis on calibration dataset, the mean  $P_2O_5$  was found to decrease from 51.68 mg/kg to 46.74 mg/kg with increasing sampling size when conditioned Latin Hypercube sampling (cLHS) was used, while it increased with increasing sampling size from 45.78 mg/kg to 47.55 mg/kg for random sampling. In a similar sequence, minimum and maximum  $P_2O_5$  except for random sampling design using 260 points decreased with increasing sampling size for both sampling strategies. However, the coefficients of variations (CVs) of  $P_2O_5$  were noted to increase with increasing sampling size for both sampling strategies. This implies that fewer points may be required for smaller uncertainty of  $P_2O_5$  prediction in the study area regardless of the sampling design to be utilized.

With an emphasis on the calibration dataset, the mean  $K_2O$  increased from 337.13 mg/kg to 348 mg/kg, increasing sampling size when cLHS was utilized. At the same time, it decreases with increasing sampling size from 357.33 mg/kg to 350.48 mg/kg with a random sampling design. Similarly, the maximum observed  $K_2O$  was also found to follow a similar trend. This implies that both sampling design and density could affect the spatial distribution of  $K_2O$ . Conversely, the coefficient of variation (CV) decreases from 81% to 78.15% with increasing sampling size when the cLHS strategy was utilized. The result

**Table 2.** Summary statistics of calibration and validation dataset P<sub>2</sub>O<sub>5</sub> (mg/kg).

Samp. Design	Dataset	Dataset	Mean	p-value	Min	Max	Variance	Std.Dev.	Coef.Var.	Skewness	Kurtosis
cLHS_10	Calibration	104	51.68	0.04	7.43	262.80	3765.20	61.36	118.74	1.92	2.61
	Validation	43	56.77	0.10	8.40	272.40	5551.83	74.51	131.25	2.09	3.37
Rand_10	Calibration	104	45.78	0.02	7.37	241.20	2892.14	53.78	117.48	2.19	4.11
	Validation	43	52.38	0.02	6.42	260.40	4096.53	64.00	122.18	1.91	2.74
cLHS_25	Calibration	260	46.74	0.07	1.65	258.00	3409.91	58.39	124.94	2.04	2.95
	Validation	108	46.16	0.01	5.91	270.00	3260.36	57.10	123.71	2.15	4.00
Rand_25	Calibration	260	47.55	0.02	3.42	271.20	3337.56	57.77	121.50	2.00	3.01
	Validation	108	42.88	0.07	5.09	217.20	2745.15	52.39	122.18	2.41	4.78

NB: Random sampling design, cLHS: conditioned Latin Hypercube sampling design, min: minimum value, max: maximum value, Std.Dev: standard deviation, Coef.Var: coefficient of variation, p-value: probability significant value for Kolmogorov–Smirnov normality test

**Table 3.** Summary statistics of calibration and validation dataset K<sub>2</sub>O (mg/kg).

Sampling design	Dataset	Dataset	Mean	p-value	Min	Max	Variance	Std.Dev.	Coef.Var.	Skewness	Kurtosis
cLHS_10	Calibration	104	337.13	0.51	0.95	1093.20	74569.51	273.07	81.00	0.45	-0.32
	Validation	43	321.49	1.00	1.83	800.40	71778.63	267.92	83.34	0.72	0.33
Rand_10	Calibration	104	357.33	0.44	1.00	1158.00	81160.17	284.89	79.73	0.45	-0.42
	Validation	43	374.02	0.37	0.89	1189.20	103550.52	321.79	86.04	0.82	0.27
cLHS_25	Calibration	260	348.00	0.35	0.89	1212.00	73965.25	271.97	78.15	0.72	0.46
	Validation	108	340.86	1.00	1.11	992.40	70143.57	264.85	77.70	0.41	-0.66
Rand_25	Calibration	260	350.48	0.54	1.11	1208.40	78917.47	280.92	80.15	0.72	0.33
	Validation	108	332.65	0.49	1.67	1030.80	62488.14	249.98	75.15	0.31	-0.52

NB: Random sampling design, cLHS: conditioned Latin Hypercube sampling design, min: minimum value, max: maximum value, Std.Dev: standard deviation, Coef.Var: coefficient of variation, p-value: probability significant value for Kolmogorov–Smirnov normality test

obtained in this study is in line with what was observed by Yu et al. (2011) in their results. They also noted the coefficients of variation (CVs) of SOC to decrease gradually from 62.8% to 47.4% with the increase in soil sampling densities. However, when using a random sampling strategy, CV was observed to increase from 79.73 to 80.15. Yu et al. (2011) established that larger CVs of the nutrient indicator are associated with more sampling points. Phosphorus showed high variability in all sample sizes and strategies in the studied soils. The result of the study implies that the detection of the variability of K<sub>2</sub>O in the study area was closely related to the sampling density and design. For instance, with the conditioned Hypercube Latin Square sampling design, the more points there are the lower level of uncertainty in the K<sub>2</sub>O spatial information acquired.

Nevertheless, the mean K<sub>2</sub>O observed in the study area was rated high, whereas P<sub>2</sub>O<sub>5</sub> was moderate. Having available soil P<sub>2</sub>O<sub>5</sub> and K<sub>2</sub>O levels adequately provides the opportunity for excellent crop yields when growing conditions are favorable. Phosphorus (P<sub>2</sub>O<sub>5</sub>) and potassium (K<sub>2</sub>O) are essential nutrients for a healthy crop. Their determination is, however, necessary when conducting soil testing.

The correlation result (Table 4) revealed that the potassium was significantly and negatively correlated with phosphorus for sampling strategies and sizes. Furthermore, the relationship was more negatively higher ( $r = -0.57$ ) with cLHS using 147 sampling points. This suggests that by increasing soil potassium, phosphorus is expected to decrease progressively and vice-versa, especially for minimum datasets. The results obtained herein are in line with previous studies. Notably, Bogunovic et al. (2017) obtained a significant and negative correlation ( $r = -0.395$ ) between phosphorus and potassium in an organic farm in Croatia. Similarly, in their studies, Hossain et al. (2014) assessed the relationship between soil pH and macronutrients in Western Nepal. They obtained a negative

**Table 4.** Relationship between P<sub>2</sub>O<sub>5</sub> and K<sub>2</sub>O, measured in mg/kg.

Sampling design and sample ratio	K <sub>2</sub> O	P <sub>2</sub> O <sub>5</sub>
clhs_10		−0.57***
rand_10		−0.55***
clhs_10		−0.55***
rand_25		−0.54***

NB: \*\*\* Significant at  $p < 0.001$

correlation coefficient ( $r = -0.028$ ) between phosphorus and potassium in the soils of Jarul forest.

### 3.2. Relative variable importance

While applying the RF models, P<sub>2</sub>O<sub>5</sub> prediction using 10 and 25% sample ratios via cLHS yielded 'y' as the relative most important variable. On the other hand, profile\_Cu was the second important variable with 10% cLHS, and bio\_12 was the second important variable with 10% cLHS. The least important variables include NDVI > TNDVI for clhs\_10 and bio\_6 > bio\_1 for clhs\_25.

Inspecting the random sampling strategies in predicting P<sub>2</sub>O<sub>5</sub>, rand\_10, and rand\_25 yielded bio\_12 and 'y' as the top two most important variables, respectively. However, in rand\_10, bio\_12 was the most important variable, while in rand\_25, 'y' was the most relevant variable. The result obtained here is different from that of Sahabiev et al. (2018) and Suleymanov et al. (2021), respectively. Their respective studies reported elevation, slope, and MMRTF (multiresolution ridge top flatness) index as the most important variables via random sampling strategy. Furthermore, the least important variable observed in rand\_10 is LRVI and TNDVI, while in rand\_25, LRVI and bio\_6 are the least important variables.

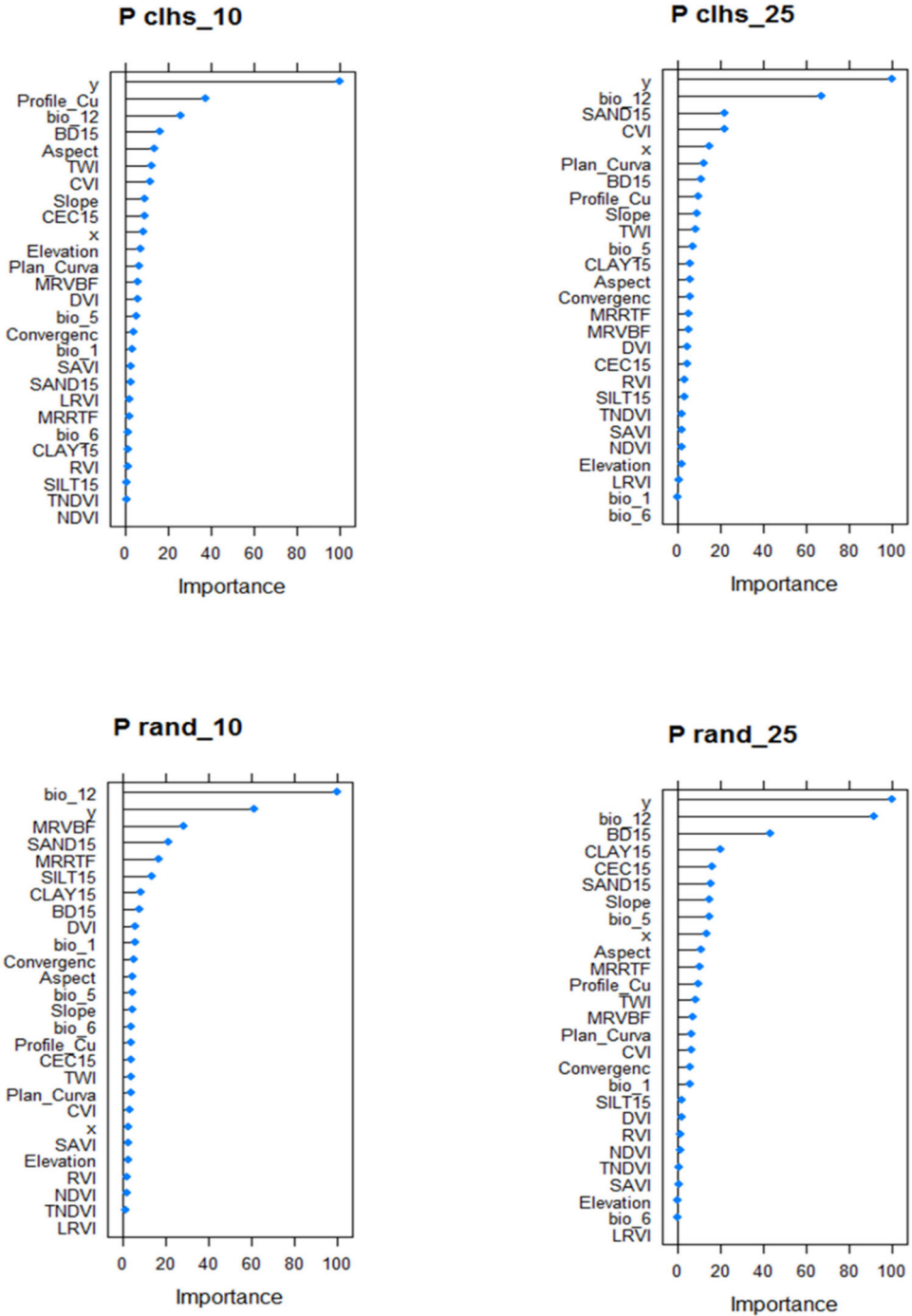
In K<sub>2</sub>O modelling, y and bio\_12 were the top two most important variables using clhs\_10 and clhs\_25, respectively. However, the least important variables varied between clhs\_10 and clhs\_25. Clhs\_10 yielded bio\_6 and clay15 as the least important variables, while clhs\_25 yielded TNDVI and RVI as the least important variables.

On the other hand, rand\_10 and rand\_25 yielded y and bio\_12 as the most important variables, respectively, while TNDV and NDVI were the least important attributes in the prediction of K<sub>2</sub>O (Figures 3 and 4).

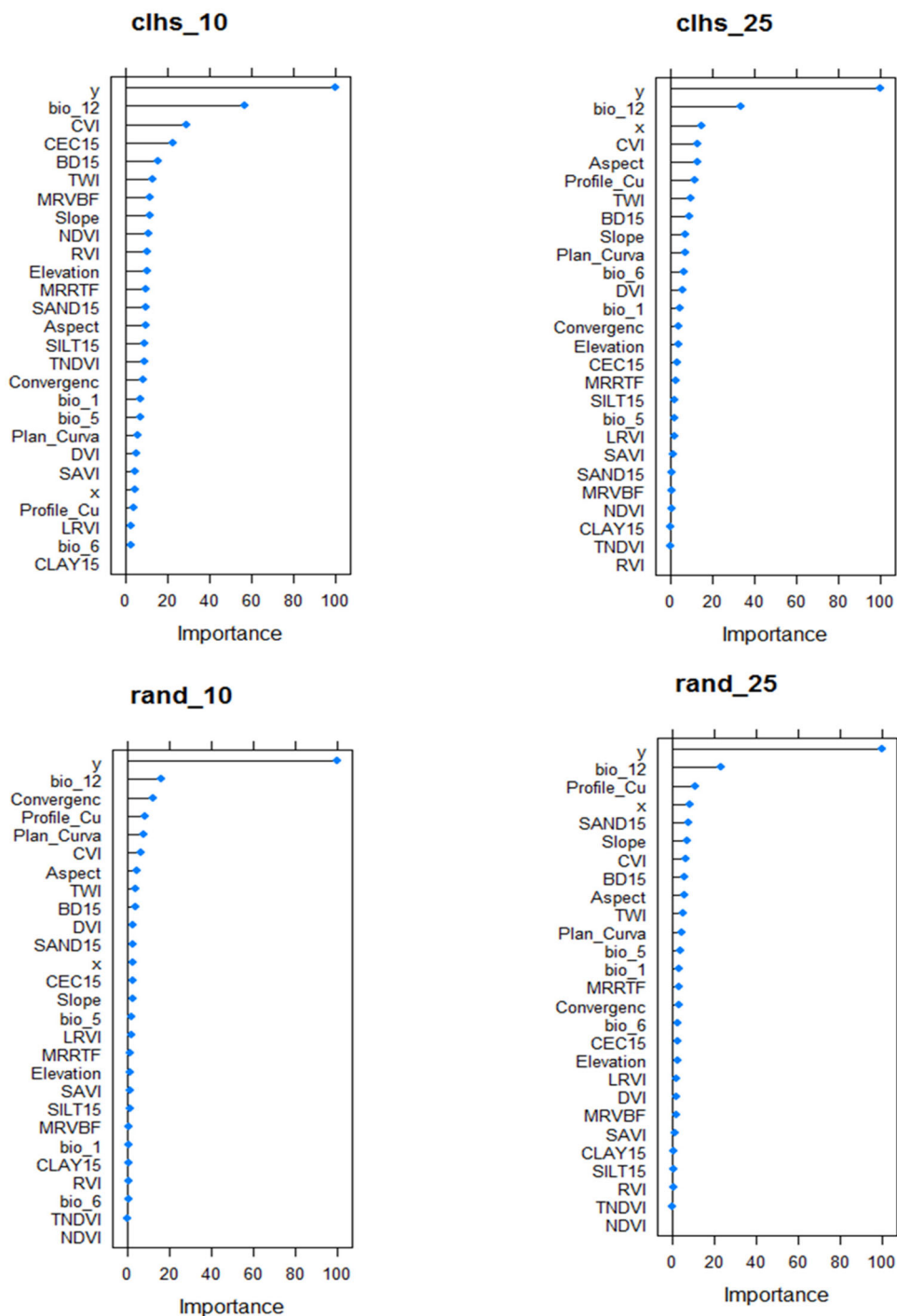
### 3.3. Model evaluation and accuracy

Table 5 presents the model evaluation and accuracy. The RMSE ranged between 42.94 – 60.74 mg/kg for P<sub>2</sub>O<sub>5</sub> prediction, R<sup>2</sup> ranged from 0.19 – 0.30, and MAE 28.87–42.59 mg/kg in the calibration dataset. Similarly, RMSE values between 29.93 – 54.88 mg/kg for phosphorus model validation, R<sup>2</sup> ranged from 0.33 – 0.82, and MAE 17.97–31.01 mg/kg for P<sub>2</sub>O<sub>5</sub> validation.

In the case of K<sub>2</sub>O prediction, when we trained our model using the calibration dataset, the model metrics, for example, RMSE, ranged from 206.64 to 252.02 mg/kg, R<sup>2</sup> (0.22–0.47), and MAE (0.35–0.47 mg/kg). However, RMSE values were between 137.77 and 181.77 mg/kg, R<sup>2</sup> (0.61–0.79), and MAE (102.46–133.38 mg/kg) in the model validation. Thus, R<sup>2</sup> values obtained in this study were within the same values obtained in other research for different soil properties (Costa et al. 2020; John et al. 2020; Lagacherie et al. 2020). In general, from these results, we can see that our model, based on most



**Figure 3.** Variable importance in P<sub>2</sub>O<sub>5</sub> (mg/kg): conditioned Latin sampling 10% sample ratio (clhs\_10) (n=147), conditioned Latin sampling 25% sample ratio (cLHS\_25) (n=368), Random sampling 10% sample ratio (rand\_10) (n=147) and D: Random sampling 25% sample ratio (rand\_25) (n=368).



**Figure 4.** Variable importance in K<sub>2</sub>O prediction (mg/kg): conditioned Latin sampling 10% sample ratio (clhs\_10) (n = 147), conditioned Latin sampling 25% sample ratio (clhs\_25) (n = 368), Random sampling 10% sample ratio (rand\_10) (n = 147) and D: Random sampling 25% sample ratio (rand\_25) (n = 368).

**Table 5.** Model evaluation and accuracy.

Soil nutrients	Dataset	Sampling design and sample ratio	RMSE	R <sup>2</sup>	MAE
P <sub>2</sub> O <sub>5</sub>	Calibration-P <sub>2</sub> O <sub>5</sub>				
	Calibration	clhs_10	60.74	0.27	42.59
	Calibration	clhs_25	52.48	0.19	35.16
	Calibration	rand_10	42.94	0.32	28.87
	Calibration	rand_25	46.95	0.31	31.42
	Validation -P <sub>2</sub> O <sub>5</sub>				
	Validation	clhs_10	54.88	0.48	31.01
	Validation	clhs_25	31.79	0.62	20.20
K <sub>2</sub> O	Validation	rand_10	43.29	0.33	24.09
	Validation	rand_25	29.93	0.82	17.97
	Calibration-K <sub>2</sub> O				
	Calibration	clhs_10	206.64	0.47	170.53
	Calibration	clhs_25	246.27	0.22	198.72
	Calibration	rand_10	252.02	0.34	207.45
	Calibration	rand_25	239.54	0.28	194.16
	Validation-K <sub>2</sub> O				
Validation	clhs_10	168.69	0.61	133.38	
Validation	clhs_25	138.02	0.79	102.46	
Validation	rand_10	181.77	0.61	128.29	
Validation	rand_25	137.77	0.74	105.42	

NB: conditioned Latin sampling 10% sample ratio (clhs\_10) (n = 147): conditioned Latin sampling 25% sample ratio (cLHS\_25) (n = 368), Random sampling 10% sample ratio (rand\_10) (n = 147) and D: Random sampling 25% sample ratio (rand\_25) (n = 368)

sampling strategies, gave a satisfactory performance in predicting soil nutrients (P<sub>2</sub>O<sub>5</sub> and K<sub>2</sub>O).

### 3.4. Role of sampling strategies in modelling phosphorus and potassium

As shown in Table 6, the accuracy metrics (RMSE, R<sup>2</sup>, and MAE) differed regarding different applied sampling strategies. For example, in cLHS, RMSE ranged from 43.34 – 153.36 mg/kg, R<sup>2</sup> ranged from 0.55 – 0.70, and MAE ranged from 25.61 – 117.92 mg/kg. In contrast, RMSE values were between 36.61 and 159.77 mg/kg in random sampling strategies, R<sup>2</sup> was from 0.58 to 0.68, and MAE was from 21.03 to 116.86 mg/kg. Furthermore, in estimating the soil nutrient elements, the random sampling strategies (i.e., rand) were more accurate in predicting P<sub>2</sub>O<sub>5</sub> (RMSE = 36.61 mg/kg, R<sup>2</sup> = 0.58 MAE = 21.03 mg/kg) than cLHS. In contrast, in the estimation of K<sub>2</sub>O, cLHS (RMSE = 153.36 mg/kg, R<sup>2</sup> = 0.70, MAE = 117.92 mg/kg) performed better than random sampling strategies. In general, accuracy metrics were better in K<sub>2</sub>O than in P<sub>2</sub>O<sub>5</sub> prediction. Also, the excellent performance of cLHS was reported by Taghizadeh-Mehrjardi et al. (2015) and Minasny and McBratney (2006). In addition, Schmidt et al. 2014 and Contreras et al. 2019 revealed that using RF in combination with cLHS gives the most accurate prediction. However, we showed that in our case, cLHS performed worse than random sampling designs in P<sub>2</sub>O<sub>5</sub> and performed better in K<sub>2</sub>O, exploiting covariates for mapping with RF.

### 3.5. Combination of sampling strategies and sample size in modelling phosphorus and potassium

Combining both sampling strategies and sample size in the spatial distribution of phosphorus and potassium for proper nutrient recommendations, rand\_25 (i.e n = 368 sample

**Table 6.** Mean of metrics of the accuracy of the studied sampling strategies (validation dataset).

Sampling strategy	Soil nutrient	RMSE	R <sup>2</sup>	MAE
cLHS	P <sub>2</sub> O <sub>5</sub>	43.34	0.55	25.61
rand		36.61	0.58	21.03
cLHS	K <sub>2</sub> O	<b>153.36</b>	<b>0.70</b>	<b>117.92</b>
rand		159.77	0.68	116.855

NB: cLHS: mean of the conditioned Latin sampling with 10% and 25% sample sizes, while rand: is Random sampling with 10% and 25% sample sizes.

size) was effective in estimating the spatial distribution of P<sub>2</sub>O<sub>5</sub> (RMSE = 29.93 mg/kg, R<sup>2</sup> = 0.82, and MAE = 17.97 mg/kg), while clhs\_25 (i.e., n = 368 sample size) gave a better K<sub>2</sub>O prediction (RMSE = 138.02 mg/kg, R<sup>2</sup> = 0.79, MAE = 102.46 mg/kg). In addition, we observed, sampling strategies with proportion increase in sample ratio gave a better prediction of the soil nutrient elements. Therefore, we can infer from the model accuracy metrics that sampling strategies with an increased sample ratio could better predict soil nutrient elements. Also, we observed that sampling strategies in prediction accuracy are susceptible to an increase in sample ratio. For example, in P<sub>2</sub>O<sub>5</sub>, RMSE decreased from 54.88 to 31.79 mg/kg, R<sup>2</sup> increased from 0.42 to 0.62, while MAE decreased from 31.01 to 20.2 in cLHS with 15% increased sample ratios. This trend was observed in both P<sub>2</sub>O<sub>5</sub> and K<sub>2</sub>O estimations, respectively. Besides that, this observation is vital for soil nutrient evaluation, and recommendation as improved soil nutrient estimates at sub-field scales are crucial for precision farming (Kozar et al. 2002).

#### 4. Conclusions

This research considered two sampling strategies and two sample ratios mapping of phosphorus and potassium, essential soil nutrients. It investigated if specific sampling strategies and sample ratios could help make precise nutrient management recommendations. The study was conducted because there was no history on the sampling strategy and sample ratios for the soil nutrient elements in the area.

The findings in the study showed that,

- Sampling strategies with increased sample ratios improved the prediction of phosphorus and potassium.
- Random sampling was suitable for phosphorus prediction, while conditioned Latin hypercubes sampling was suitable for potassium.
- Conditioned Latin hypercubes sampling was the overall best sampling strategy

In conclusion, the adopted approach showed the prospect of precisely and accurately predicting soil nutrients in the Mediterranean region. Furthermore, the conditioned Latin hypercubes sampling strategy shows success and robustness in predicting soil nutrients. The above findings are not only relevant for the investigated area. However, similar investigations remain needed in other fields, agroecosystems, and climatic conditions before their generalization. Nevertheless, it is highly recommended that conditioned Latin hypercubes sampling with increased sample size be adopted for further soil nutrient evaluation studies and to support management decision-making.

#### Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Yassine Bouslihim  <http://orcid.org/0000-0002-3666-1850>  
 Abdelkrim Bouasria  <http://orcid.org/0000-0002-9101-6520>  
 Rachid Razouk  <http://orcid.org/0000-0003-0855-3462>  
 Lahcen Hssaini  <http://orcid.org/0000-0002-6739-3895>

## Data availability statement

The data supporting this study's findings are available from the corresponding author, [JK], upon reasonable request.

## References

- Abbasi MK, Majeed A, Sadiq A, Khan SR. 2008. Application of Bradyrhizobium japonicum and phosphorus fertilization improved growth, yield and nodulation of soybean in the sub-humid hilly region of Azad Jammu and Kashmir, Pakistan. *Plant Prod Sci.* 11(3):368–376.
- Allali A, Rezouki S, Lougramzi H, Touati N, Eloutassi N, Fadli M. 2020. Agricultural traditional practices and risks of using insecticides during seed storage in Morocco. *Plant Cell Biotechnol Mol Biol.* 21(39–40):29–37.
- Bogunovic I, Pereira P, Brevik EC. 2017. Spatial distribution of soil chemical properties in an organic farm in Croatia. *Sci Total Environ.* 584–585:535–545.
- Bouslihim Y, Rochdi A, Aboutayeb R, El Amrani-Paaza N, Miftah A, Hssaini L. 2021. Soil aggregate stability mapping using remote sensing and GIS-based machine learning technique. *Front Earth Sci.* 9: 748859.
- Breiman L. 2001. Random forests. *Mach Learn.* 45(1):5–32.
- Brungard CW, Boettinger JL. 2010. Conditioned Latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors. *Digital soil mapping. Progress in soil science, vol 2.* Dordrecht: Springer.
- Brus DJ, De Gruijter JJ. 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma.* 80(1-2):1–44.
- Contreras J, Ballari D, De Bruin S, Samaniego E. 2019. Rainfall monitoring network design using conditioned Latin hypercube sampling and satellite precipitation estimates: an application in the ungauged Ecuadorian Amazon. *Int J Climatol.* 39(4):2209–2226.
- Costa EM, Pinheiro HSK, Anjos LHCD, Marcondes RAT, Gelsleichter YA. 2020. Mapping soil properties in a poorly-accessible area. *Rev Bras Ciênc do Solo.* 44:e0190107
- FAO, ISRIC and ISSS. 2006. World Reference Base for Soil Resources 2006, A framework for international classification, correlation and communication. FAO, World Soil Res Rep. 103:1-128, Rome.
- Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol.* 37(12):4302–4315.
- Gumiere T, Rousseau AN, da Costa DP, Cassetari A, Cotta SR, Andreote FD, Gumiere SJ, Pavinato PS. 2019. Phosphorus source driving the soil microbial interactions and improving sugarcane development. *Sci Rep.* 9(1):1–9.
- Higo M, Isobe K, Yamaguchi M, Toriogoe Y. 2015. Impact of a soil sampling strategy on the spatial distribution and diversity of arbuscular mycorrhizal communities at a small scale in two winter cover crop rotational systems. *Ann Microbiol.* 65(2):985–993.
- Hossain I, Osman KT, Kashem A, Sarker A. 2014. Correlations of available phosphorus and potassium with pH and organic matter content in the different forested soils of Chittagong Hill Tracts, Bangladesh. *Int J Forest Soil Erosion.* 4(1):7–10.
- Hyndman RJ, Athanasopoulos G. 2018. *Forecasting: principles and practice.* Melbourne: OTexts. <https://otexts.com/fpp2/>.
- Jeong G, Choi K, Spohn M, Park SJ, Huwe B, Liefß M. 2017. Environmental drivers of spatial patterns of topsoil nitrogen and phosphorus under monsoon conditions in a complex terrain of South Korea. *PLoS One.* 12(8):e0183205.
- John K, Abraham Isong I, Michael Kebonye N, Okon Ayito E, Chapman Agyeman P, Marcus Afu S. 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land.* 9(12):487.



- Kadarwati TF. 2020. Effect of different levels of potassium on the growth and yield of sugarcane ratoon in inceptisols. In: IOP Conference Series: Earth and Environmental Science. Vol. 418; p. 012066. IOP Publishing.
- Kozar B, Lawrence R, Long DS. 2002. Soil phosphorus and potassium mapping using a spatial correlation model incorporating terrain slope gradient. *Precis Agric.* 3(4):407–417.
- Kradi C. 2012. L'agriculture solidaire dans les écosystèmes fragiles au Maroc. Morocco: INRA.
- Lagacherie P, Arrouays D, Bourennane H, Gomez C, Nkuba-Kasanda L. 2020. Analyzing the impact of soil spatial sampling on the performances of digital soil mapping models and their evaluation: a numerical experiment on quantile random forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. *Geoderma.* 375:114503.
- Malone BP, Minansy B, Brungard C. 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. *PeerJ.* 7:e6451.
- Marschner H, Dell B. 1994. Nutrient uptake in mycorrhizal symbiosis. *Plant Soil.* 159(1):89–102.
- Mesrar H, Sadiki A, Faleh A, Quijano L, Gaspar L, Navas A. 2017. Vertical and lateral distribution of fall-out <sup>137</sup>Cs and soil properties along representative toposequences of central Rif, Morocco. *J Environ Radioact.* 169-170:27–39.
- Minansy B, McBratney AB. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Comput Geosci.* 32(9):1378–1388.
- Mirzaee S, Ghorbani-Dashtaki S, Mohammadi J, Asadi H, Asadzadeh F. 2016. Spatial variability of soil organic matter using remote sensing data. *Catena.* 145:118–127.
- Morgan JÁ, Connolly EÁ. 2013. Plant-soil interactions: nutrient uptake. *Nat Educ Knowl.* 4(8):2.
- Öborn I, Andrist-Rangel Y, Askegaard M, Grant CA, Watson CA, Edwards AC. 2005. Critical aspects of potassium management in agricultural systems. *Soil Use Manage.* 21(s1):102–112.
- Oosterhuis DM, Loka DA, Kawakami EM, Pettigrew WT. 2014. The physiology of potassium in crop production. *Adv Agron.* 126:203–233.
- Ramaekers L, Remans R, Rao IM, Blair MW, Vanderleyden J. 2010. Strategies for improving phosphorus acquisition efficiency of crop plants. *Field Crops Res.* 117(2–3):169–176.
- Rezouki S, Allali A, Louasté B, Eloutassi N, Fadli M. 2020. Physico-chemical evaluation of soil resources in different regions of Taza-Taounate. *Mediterr J Chem.* 10(9):836. 11(1), pp.1
- Sahabiev IA, Ryazanov SS, Kolcova TG, Grigoryan BR. 2018. Selection of a geostatistical method to interpolate soil properties of the state crop testing fields using attributes of a digital terrain model. *Eurasian Soil Sci.* 51(3):255–267.
- Scarpone C, Schmidt MG, Bulmer CE, Knudby A. 2016. Modelling soil thickness in the critical zone for Southern British Columbia. *Geoderma.* 282:59–69.
- Schmidt K, Behrens T, Daumann J, Ramirez-Lopez L, Werban U, Dietrich P, Scholten T. 2014. A comparison of calibration sampling schemes at the field scale. *Geoderma.* 232:243–256.
- Suleymanov A, Abakumov E, Suleymanov R, Gabbasova I, Komissarov M. 2021. The soil nutrient digital mapping for precision agriculture cases in the trans-ural steppe zone of Russia using topographic attributes. *ISPRS Int J Geo-Inf.* 10(4):243.
- Sun XL, Wang HL, Zhao YG, Zhang C, Zhang GL. 2017. Digital soil mapping based on wavelet decomposed components of environmental covariates. *Geoderma.* 303:118–132.
- Taghizadeh-Mehrjardi R, Sarmadian F, Tazeh M, Omid M, Toomanian N, Rosta MJ. 2015. Comparison of different Sampling methods for digital soil mapping in Ardakan region. *J Manage Watershed Eng.* 4(6):353–363.
- Thomas M, Clifford D, Bartley R, Philip S, Brough D, Gregory L, Willis R, Glover M. 2015. Putting regional digital soil mapping into practice in Tropical Northern Australia. *Geoderma.* 241–242: 145–157. 2015
- Varvel GE, Schlemmer MR, Schepers JS. 1999. Relationship between spectral data from an aerial image and soil organic matter and phosphorus levels. *Precis Agric.* 1(3):291–300.
- Wollenhaupt NC, Wolkowski RP, Clayton MK. 1994. Mapping soil test phosphorus and potassium for variable-rate fertilizer application. *J Prod Agri.* 7(4):441–448.
- Yu D, Zhang Z, Yang H, Shi X, Tan M, Sun W, Wang H. 2011. Effect of soil sampling density on detected spatial variability of soil organic carbon in a red soil region of China. *Pedosphere.* 21(2): 207–213.