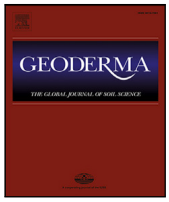
Geoderma 441 (2024) 116740

Contents lists available at ScienceDirect

Geoderma

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



Evaluating the extrapolation potential of random forest digital soil mapping

Fatemeh Hateffard [a], Luc Steinbuch [b,*], Gerard B.M. Heuvelink [b,c]

[a] Department of Landscape Protection and Environmental Geography, University of Debrecen, Egyetem ter 1, H-4032 Debrecen, Hungary
[b] Soil Geography and Landscape group, Wageningen University and Research, Wageningen, The Netherlands
[c] ISRIC – World Soil Information, Wageningen, The Netherlands

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Spatial soil information
Extrapolation effects
Prediction accuracy
Similarities

ABSTRACT

Spatial soil information is essential for informed decision-making in a wide range of fields. Digital soil mapping (DSM) using machine learning algorithms has become a popular approach for generating soil maps. DSM capitalises on the relation between environmental variables (i.e., features) and a soil property of interest. It typically needs a training dataset that covers the feature space well. Mapping in areas where there are no training data is challenging, because extrapolation in geographic space often induces extrapolation in feature space and can seriously deteriorate prediction accuracy. The objective of this study was to analyse the extrapolation effects of random forest DSM models by predicting topsoil properties (OC, clay, and pH) in four African countries using soil data from the ISRIC Africa Soil Profiles database. The study was conducted in eight experiments whereby soil data from one or three countries were used to predict in the other countries. We calculated similarities between donor and recipient areas using four measures, including soil type similarity, homosoil, dissimilarity index by area of applicability (AOA), and quantile regression forest (QRF) prediction interval width. The aim was to determine the level of agreement between these four measures and identify the method that had the strongest agreement with common validation metrics. The results indicated a positive correlation between soil type similarity, homosoil and dissimilarity index by AOA. Surprisingly, we observed a negative correlation between dissimilarity index by AOA and QRF prediction interval width. Although the cross-validation results for the trained models were acceptable, the extrapolation results were unsatisfactory, highlighting the risk of extrapolation. Using soil data from three countries instead of one increased the similarities for all measures, but it had a limited effect on improving extrapolation. Also, none of the measures had a strong correlation with the validation metrics. This was particularly disappointing for AOA and QRF, which we had expected to be strong indicators of extrapolation prediction performance. Results showed that homosoil and soil type methods had the strongest correlation with validation metrics. The results for this case study revealed limitations of using AOA and QRF as measures of extrapolation effects, highlighting the importance of not relying on these methods blindly. Further research and more case studies are needed to address the effects of extrapolation of DSM models.

1. Introduction

Spatial soil information in the form of maps is essential in detailed soil quality assessments, sustainable land management, and precision agriculture studies (Lagacherie and McBratney, 2006). Nowadays soil maps are most often made by digital soil mapping (DSM), where machine learning (ML) is a frequently used mapping algorithm. Machine learning first captures the relation between environmental variables and the soil property of interest using training data and next uses this relation to spatially predict the soil property from maps of the environmental variables (McBratney et al., 2003). Advances in remote sensing provide ever increasing spatial and detailed information of environmental variables (Yang et al., 2011; Asgari et al., 2020). The successful application of a data-driven technique such as ML also requires fairly large training datasets. Moreover, the training data should cover the feature space well, meaning that ranges and combinations of environmental variables present in the study area are adequately represented in the training data set (Minasny and McBratney, 2010; Ng et al., 2018; Wadoux et al., 2019; Hateffard and Novák, 2021). The choice of ML algorithm also matters since the algorithm should be able to learn the complex relationship between environmental covariates and soil properties from the data. Among different ML techniques, random forest has proven its applicability in spatial prediction of soil properties in several studies (Ließ et al., 2012; Vaysse and Lagacherie, 2015; Kinoshita et al., 2016; Hengl et al., 2018).

* Corresponding author.
E-mail address: luc.steinbuch@wur.nl (L. Steinbuch).

https://doi.org/10.1016/j.geoderma.2023.116740
Received 9 July 2023; Received in revised form 12 October 2023; Accepted 1 December 2023
Available online 23 December 2023

In practice, field surveys, soil sampling and laboratory analyses are expensive; therefore often legacy soil data are used in DSM studies (Tan, 1995; Arrouays et al., 2020). The sampling density can vary strongly between regions and large parts of the study area might not be represented in the training data or have low sampling density (Minasny et al., 2020). Mapping in such areas is challenging when there are no resources to collect new soil samples. In such cases spatial extrapolation, i.e. using soil data from one area to predict in another area, might be a potential solution. But extrapolation can amplify prediction uncertainty and should ideally be applied in areas with similar soil forming factors. One might expect that, soils with similar soil-forming factors will likely have the same soil conditions (Jenny, 1994).

Spatial extrapolation likely works well if a model is developed with data from an area that has good coverage of the soil forming factors (Afshar et al., 2018; Neyestani et al., 2021), but in practice the training data from one area might not cover the feature space of another area well. In other words, extrapolation in geographical space might lead to extrapolation in feature space. If a ML model is employed where the feature space between the two areas differs considerably, it may produce inaccurate and unreliable predictions (Meyer and Pebesma, 2021). This is particularly relevant in case of continental and global mapping of soil properties (e.g. Arrouays et al. (2014), Batjes et al. (2020) and Poggio et al. (2021)). These considerations have led to the development of the concept of "Area of Applicability (AOA)" (Meyer and Pebesma, 2021), which calculates a dissimilarity index between covariates in the training data and covariates at prediction locations and delineates the area where extrapolation in feature space occurs. Based on AOA, we should only predict in regions that have similar conditions as the area seen by the model.

Apart from AOA, there are also other metrics to investigate the degree of extrapolation in DSM. For example, Mallavan et al. (2010) introduced the homosoil method as a helpful way to decide which areas have similar soils as a source area. As long as the source area sufficiently captures the environmental heterogeneity and the soil-forming factors are similar to those in the prediction area, a model trained in the source area is judged useful for extrapolation (Bui and Moran, 2003; Nenkam et al., 2022). Alternatively, taking into account that the soil conditions are summarised by soil type, comparison of soil type maps between the source and prediction area is also informative about the extrapolation potential (Angelini et al., 2020). These methods, homosoil and soil types, are alternative tools to AOA to evaluate whether extrapolation in geographic space is feasible.

If extrapolation in geographic space leads to extrapolation in feature space then this will likely also show up in the prediction uncertainty as quantified by some ML methods. Uncertainty estimation of soil maps through quantile regression forests (QRF) (Meinshausen and Ridgeway, 2006) provides quantiles of the conditional distribution from which prediction intervals can be derived. Thus a map of the prediction interval width (PIW) can be produced as a by-product of QRF, by subtracting the lower from the upper quantile for any point in the area of interest (Zhang et al., 2019). Areas where the PIW is larger than a threshold could be considered too uncertain to be mapped (Vaysse and Lagacherie, 2017). It would be interesting to evaluate to what degree these areas overlap with extrapolation areas identified by the AOA method. If the two methods have strong agreement, then QRF might be an easier way to evaluate which areas can and cannot be predicted using a model that was trained in a specific area.

In previous studies, different researchers have applied different extrapolation methods between two similar areas for mapping soil classes and properties (Grinand et al., 2008; Malone et al., 2016; Zhang et al., 2018; Du et al., 2021). Malone et al. (2016) evaluated the similarity of the environment between the donor and recipient areas utilising the homosoil approach by quantifying a taxonomic distance measure and then extrapolated the model from one region to another. Afshar et al. (2018) investigated the similarity index between two areas by Gower's similarity index and applied a multinomial logistic regression model to estimate soil great groups. They found that the extrapolation was successful within the recipient area up to 60% prediction accuracy. Angelini et al. (2020) applied Structural Equation Modelling as a technique that includes expert knowledge to analyse the capability to extrapolate a model from one area to another. They concluded that quantifying all soil-environment interactions over time is still challenging, and that we need a better understanding of these aspects. Nenkam et al. (2022) challenged the possibility of extrapolation in areas assumed to be similar based on the homosoil approach, and compared the results with existing global maps. They found that extrapolation in geographic space is feasible, however the accuracy can be improved if local data are included in the training dataset.

The review above shows that there are many different ways to determine the potential of extrapolating DSM models trained in one area to other areas. These methods include homosoil, soil type similarity, dissimilarity index by AOA, and QRF prediction interval width. The objective of this study was to investigate which method has the strongest agreement with statistical validation metrics computed from data in the prediction area. Based on such analysis we aimed to gain insight into which similarity metrics are the best indicators of whether spatial extrapolation occurs and leads to poorer prediction performance.

To achieve the objective, we: (1) estimated the similarity of soil forming factors between donor and recipient areas by using the soil types and homosoil approaches; (2) trained a RF model on data from a donor area, extrapolated it to a recipient area, and computed dissimilarity index by AOA and QRF prediction interval width; and (3) evaluated the agreement between the four "measures of similarity" and common statistical validation metrics, computed using independent data from the recipient area.

We performed the tasks above by means of a case study. We selected four African countries and used data from the ISRIC Africa Soil Profiles (AfSP) database (Leenaars et al., 2013) to train DSM models and evaluate their performance, using different combinations of countries as donor and recipient areas. For reasons explained later, we consider organic carbon (OC) content, clay content and pH as the soil properties of interest.

2. Materials and methods

2.1. Study area

We selected four African countries as our study area: Ethiopia, Kenya, Burkina Faso, and Nigeria (Fig. 1). The reasons for selecting these countries were twofold: first, we wanted similar and dissimilar countries to assess different degrees of extrapolation; second, we required that there were sufficient soil samples in a public database for each country and that the data had a fairly uniform spatial distribution across each country. Kenya and Ethiopia are located in the same region in North-East Africa and share comparable climates with hot arid lowlands and cool moist highlands. In Kenya, the climatic conditions range from humid in the west to arid in the east and north. In Ethiopia, the southeast and northeast regions have a warm desert climate, primarily in the lowlands, while the central and western highlands have a humid subtropical and tropical savanna climate. Nigeria and Burkina Faso have similarities in terms of climate, with hot and humid tropical conditions in the south of Nigeria, sub-humid savanna conditions in the south of Burkina Faso and arid and semi-arid conditions in the north. The far northern parts of both countries are mainly desert areas with sparse vegetation. Humidity increases southwards together with more abundant vegetation. Apart from these similarities, each country also experiences its own specific climate since Nigeria has also coastal conditions (tropical monsoon climate)(https://climateknowledgeportal.worldbank.org/).

In terms of topography, in Ethiopia most of the country is covered by the Ethiopian Highlands which are characterised by undulating plateaus dissected by steep slopes and deep valleys. The lowlands of

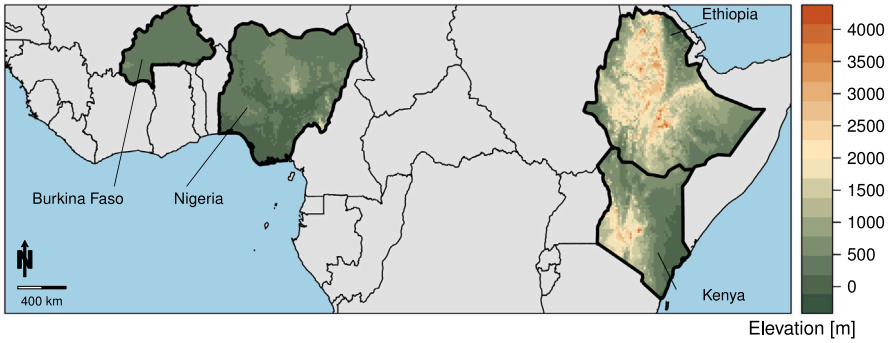


Fig. 1. Location and digital elevation model of four African countries that together form the study area.

Ethiopia are located in the east and southeast and a narrow strip through the centre. Elevation differences in Ethiopia are large, with peaks up to 4411 m and a lowest point at 125 m below sea level. Kenya has several mountains and large plateaus as well as large lowland plains. The central parts of Nigeria are dominated by rolling hills and high plains while the country's northern regions are characterised by relatively flat plains. Burkina Faso has a relatively flat, slightly undulating, landscape where the maximum elevation difference is about 700 m (Jones et al., 2013).

Regarding soil types (Panagos et al., 2012), Kenya has the largest soil diversity, with fertile volcanic soils in the western highlands, and sandy and rocky soils dominating the eastern lowlands. In Ethiopia, around one-third of the soils are shallow over hard bedrock, especially in the mountainous parts in the north and most of the lowlands in the east. Fertile to very fertile soils can be found in much of the highlands, which are though exposed to erosion, including Luvisols and Nitisols as well as Vertisols, which are less well drained and less suitable for cultivation. In the north and centre of Nigeria, easily erodible sandy and loamy soils of low fertility occur, like Arenosols and Lixisols, while in the south there are deep red clayey soils with a well-developed structure and high productivity (Nitisols). Burkina Faso has less variety in soil types compared to the other three countries. In general, the soils in Burkina Faso are loamy, gravelly and often shallow, with low fertility (Lixisols and Plinthosols). In the northern parts of the country, the soils are exposed to degradation and desertification. Nevertheless, some parts of Burkina Faso have fertile clayey soils that are suitable for agriculture (Luvisols), such as on the foot slopes of metamorphic hills and the plains near the main rivers (see Table SM-1 in the Supplementary Materials for an overview of soil types per country).

Kenya and Ethiopia have a diverse land cover which includes cropland, shrubland, grassland, and forests on complex terrain. The highlands in Ethiopia are covered by forests and grasslands, while arid and semi-arid areas in the lowlands are covered by scrub vegetation or are bare. The land cover in Kenya is largely covered by savannas characterised by grasslands mixed with scattered trees. The main land cover types in Nigeria and Burkina Faso are shrubland and grassland, and forests in the south of Nigeria, with croplands and grazing lands mainly occurring on the relatively lower parts of the undulating landscapes.

2.2. Soil data and covariates

The ISRIC Africa Soil Profiles (AfSP) database (Leenaars et al., 2014) includes a compilation of nearly 18,000 soil profiles from various digital and analogue data sources covering most parts of Africa. We chose pH-H2O, Organic Carbon (OC) content, and Clay content as the target soil properties for modelling and mapping. These are important soil properties and had a sufficiently large sample size for all four countries (Table 1). As we only focus on topsoil characteristics, the selected depth interval was 0 to 20 cm. Since the AfSP observations contain different depth intervals, the observations were harmonised by taking a weighted average if there were multiple observed layers within the 0–20 cm depth interval. If less than 15 cm of the selected depth interval was covered by the observations on a location, that location was ignored. Summary statistics of the three soil properties are given in Table 1.

Table 1
Summary statistics of AfSP observations of three soil properties for the four countries. SD is standard deviation and *n* is number of observations. The unit for Clay is g/100 g and for OC g kg^{-1}.

Country	Variable	Min	Mean	Max	SD	*n*
Kenya	Clay	0.0	37.7	88.0	19.6	400
	OC	0.3	14.4	360.0	19.2	848
	pH	4.0	7.0	11.0	1.2	845
Ethiopia	Clay	2.0	35.4	90.0	17.3	1082
	OC	0.6	24.4	251.0	23.6	1661
	pH	4.0	7.0	9.9	1.1	1710
Nigeria	Clay	0.0	20.2	84.0	18.9	1074
	OC	0.2	9.7	102.4	8.3	1667
	pH	3.0	6.0	9.3	0.8	1753
Burkina Faso	Clay	1.0	21.5	64.0	14.8	616
	OC	0.9	9.4	43.1	6.4	613
	pH	4.6	6.4	8.8	0.6	595

A set of 35 environmental covariates that represent soil forming factors was used, including covariates representing climate conditions, topography, and vegetation (Table SM-2 in Supplementary Materials). Also, from the Digital Elevation Model, which is the primary representation of topography, 14 covariates were extracted using the RSAGA package (Brenning et al., 2018) in R (R Core Team et al., 2021). All covariates were resampled to a 1 km spatial resolution.

2.3. Experimental set-up

To investigate the effects of extrapolation, we used the following set-up: (1) train the model on all data from each country individually, and predict to that country and the other three countries; (2) train the model on data from three countries, and predict to these three countries and the fourth. Thus in total, we had eight models for each of the three soil properties. A country, or a combination of countries, that is used for calibration is indicated as the "donor" area; a country or countries that is extrapolated into is indicated as a "recipient" area.

Predictions of target soil properties were based on the random forest algorithm. This model was calibrated using the *caret* package. Random forest (RF; Breiman, 1996) fits many decision trees (independent from each other) with a random sample of covariates chosen at each splitting node. In this study, we chose to stick with the default values of hyperparameters for the RF model in our experiments. Cross-validation was employed to evaluate and compare the performance of each RF model for donor areas. Here, we applied 10-fold cross-validation. In

this validation method, the dataset is randomly divided into ten folds of similar size, and each time, one of the folds is kept aside and used for validation of predictions made using calibration data from the other nine folds. This procedure was repeated ten times so that each fold was utilised exactly once for validation. Next, the mean error (ME), root mean square error (RMSE) and model efficiency coefficient (MEC) (Nash and Sutcliffe, 1970) accuracy metrics were computed:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \tag{2}$$

$$MEC = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{3}$$

where n is the number of observations, O_i is the observed value for the ith location; P_i is the predicted value for the ith location; and $\overline{O}$ the mean of the observations. In recipient areas we used all data from that area for validation, since none of them were used for model calibration. For both the calibration and validation of each model, we utilised all available data, without selectively choosing points that fall within similar areas.

2.4. Measures of extrapolation

We used four methods to characterise the degree of extrapolation, as described in the four subsections below.

2.4.1. Similarity in soil types

Based on the Soil Atlas of Africa (Panagos et al., 2012), which contains the dominant WRB reference soil types represented as spatial polygons, we first calculated the percentage of each soil type in each country. Next we assessed the similarity between the soil types of two countries using the Jaccard measure of similarity (Awad and Khanna, 2015, page 36) by accumulating the minimum percentages of each combination of two countries sharing the same soil type:

$$Sim_{ij} = \sum_{k \in K} min(A_{ik}, A_{jk}) \tag{4}$$

with Sim_{ij} the Jaccard similarity measure between countries i and j; K all soil types and A_{ik} and A_{jk} the proportion of area of soil type k in countries i and j, respectively. The Jaccard similarity is a number between 0 and 1, where a value of 0 means no similarity and a value of 1 means perfect similarity.

Additionally, we calculated the Jaccard measure of similarity taking also the taxonomic distance between soil types into account. The taxonomic distance quantifies the degree of similarity between soil types. We used the taxonomic distances specified in Minasny et al. (2010), which first assigns 21 binary key features such as "calcareous" or "accumulation of silica" to each soil type, and next computes the Euclidean distance between all soil types in the resulting – 21-dimensional – key feature space. Finally, these relative distances are scaled to values between 0 and 1 to obtain the taxonomic distance. For every soil type combination, except for any soil type with itself, we calculated the smallest shared proportion between two countries as also done for the Jaccard similarity, and next we multiplied this proportion with $1 - taxonomic\ distance$ (because we want to express the similarity, not the dissimilarity). We added the results for all combinations and divided the outcome by the the same sum in the theoretical situation that all soil type combinations have zero taxonomic distance (i.e., when there is maximal similarity between all soil types):

$$a_{ij} = \frac{\sum_{m \in K}\sum_{n \in K,\ n \neq m} min(A_{im},\ A_{jn}) \cdot (1 - TD_{mn})}{\sum_{m \in K}\sum_{n \in K,\ n \neq m} min(A_{im},\ A_{jn})} \tag{5}$$

with a_{ij} an addition factor for the similarity measure based on taxonomic distance (a number between 0 and 1, representing minimal and maximal additional similarity respectively) and TD_{mn} the taxonomic distance between soil types m and n.

Finally, the outcome of Eq. (5) was scaled between Sim_{ij} and 1 before being added to Sim_{ij}, in such a way that if all taxonomic distances would be maximal, the scaled similarity with taxonomic distance $SimTDsc_{ij}$ would equal Sim_{ij}, and if all taxonomic distances would be minimal, $SimTDsc_{ij}$ would equal one:

$$SimTDsc_{ij} = Sim_{ij} + a_{ij}(1 - Sim_{ij}). \tag{6}$$

2.4.2. Homosoil fraction

An alternative approach to quantify similarity between areas is the homosoil approach, developed by Mallavan et al. (2010). The underlying theory of this method is based on the taxonomic distance (Booth et al., 1987) of the environmental covariates between the donor and recipient areas, where these covariates represent key soil-forming factors. Mallavan et al. (2010) created a spatial database of environmental variables at the global scale, including climate, topography, and lithology/parent material. The method calculates Gower's similarity index at three hierarchical levels, first by selecting the areas with similar climate conditions (homoclime), then choosing the same lithological classes within homoclime areas (homolith), and last by deriving the similar topography (homotop) in previously selected homoclime and homolith areas. We applied this method to identify similarity in terms of soil forming factors between locations in the donor and recipient countries. The assumption is that if the soil forming factors are similar, the two locations are "homosoil" and have similar soils. In this study, for each donor pixel, we calculated a map layer of the recipient country indicating the homosoil pixels. Those map layers are combined into one final map where each pixel indicates if it is homosoil in at least one of the map layers. From this final map, the "homosoil similarity" was calculated as the fraction of the surface in the recipient area which is homosoil to at least one location (or grid cell) in the donor area. Note that unlike the soil type similarities, the homosoil fraction is asymmetric, as are the similarity measures discussed in the next subsections.

2.4.3. Dissimilarity index by AOA

Area of Applicability (AOA) is a solution to prevent extrapolation issues in machine learning models proposed by Meyer and Pebesma (2021). It limits predictions to areas where the covariates are similar to the covariates at training locations. It works by first computing a dissimilarity index between donor and recipient locations using distances in covariate space between the two locations, and weighting covariates according to their importance in the machine learning model , trained on all data from the donor area. Next a threshold is applied whereby all prediction locations with a dissimilarity index below the threshold are assigned to the AOA. The AOA function, which is implemented in the CAST package in R (Meyer et al., 2023), has two output layers: the dissimilarity index (DI) and the area of applicability (AOA). DI can take any value between 0 and infinity, where larger values indicate a larger dissimilarity. The AOA layer has only two values, 0 and 1, where 1 indicates that a location belongs to the AOA, and 0 that it does not. In this study, we only used the DI layer. To speed up calculations, we reduced the resolution of the covariate data before applying of the AOA with a factor of 10.

2.4.4. QRF prediction interval width

Finally, we also used QRF to compute the width of 90% prediction intervals (Meinshausen and Ridgeway, 2006) at all prediction locations in a recipient area, when using a model trained on data from the donor area. The 90% PIW is calculated by subtracting the 0.05 quantile from the 0.95 quantile. In case of extrapolation it is expected that the prediction intervals are wider. To speed up calculations we reduced the resolution of the covariate data, as previously done in Section 2.4.3.

Table 2
Similarity between countries, considering soil types. The lower left triangle presents plain similarity; the upper right triangle similarity while taking taxonomic distance into account. The values around the diagonal show the soil type similarity for a recipient country when the three other countries are a combined donor.

	Kenya	Ethiopia	Nigeria	Burkina Faso
Kenya	50.3%	86.6% 58.9%	61.4%	48.2%
Ethiopia	38.0%	39.8%	81.8% 55.0%	60.1%
Nigeria	41.3%	35.3%	48.0%	75.9% 65.0%
Burkina Faso	26.0%	43.4%	46.5%	46.7% 80.5%

Table 3
Homosoil scores. Columns are donor countries, rows recipient countries. The diagonal values (shown in boxes) shows the homosoil score for a recipient country when all three other countries are a donor.

	Kenya	Ethiopia	Nigeria	Burkina Faso
Kenya	56.3%	29.5%	33.2%	0.5%
Ethiopia	41.4%	48.5%	14.2%	1.1%
Nigeria	36.5%	14.6%	45.5%	14.0%
Burkina Faso	6.6%	20.9%	14.3%	30.8%

3. Results

3.1. Similarities in soil types

Before fitting and applying the random forest models and evaluating the extrapolation potential using AOA and QRF prediction interval widths, similarities between the four countries and their combinations were first checked in terms of soil types and homosoil. Table 2 presents the similarities regarding soil types in the four countries. Burkina Faso and Nigeria had the highest soil type similarity, both for plain Jaccard similarity and for similarity that accounts for taxonomic distance. The lowest similarity was obtained between Burkina Faso and Kenya, with a value of 26.0%. Countries tend to have more similar soil types if they are from the same region (West or East Africa), although Kenya and Nigeria also have fairly high similarities. Incorporating taxonomic distance slightly increased the similarity between countries, which is due to different soil class combinations having a taxonomic distance smaller than the maximum.

Generally speaking, "similarity while taking taxonomic distance into account" was 20% higher than "plain similarity". The biggest difference between plain similarity and similarity accounting for taxonomic distance was for Burkina Faso and Ethiopia, which indicates that these countries benefit most from sharing common factors that influence soil formation.

Combining soil type from three countries, the highest similarity in soil types was observed when Kenya was the recipient country, for both plain similarity and considering taxonomic distance. Remarkably, the inclusion of soil type data from three countries and the incorporation of taxonomic distance into the similarity calculation resulted in considerably higher values of 75% to 86% compared to experiments that relied on plain similarity or data from a single country.

3.2. Homosoil

The homosoil method assesses similarity between areas based on their similarity of soil forming factors. Table 3 shows the homosoil scores. Recall that we calculated the homosoil scores in two ways: (1) one country is the donor and all other countries are recipients; (2) three countries are the donor and the fourth country is a recipient. When Kenya is the donor country, the homosoil scores for Ethiopia (41%) and Nigeria (36%) are high, while the score for Burkina Faso is low. According to the homosoil concept Burkina Faso is quite different from the other three countries, because all have low homosoil scores if Burkina Faso is the donor country. Table 3 also shows as expected that having three countries as a donor increases the possibility of finding more similar soil forming factors in the recipient country. This is shown by the higher homosoil scores. Note, however, that combination of three countries to find similar soils in Burkina Faso still has a low score, lower than when Kenya is a single donor of Ethiopia and Nigeria and when Nigeria is a donor of Kenya.

3.3. Machine learning model and dissimilarity index by AOA

To be able to compute the dissimilarity index and AOA, we first needed to train a random forest model for each experiment and soil property. The performance of the random forest model with default hyperparameter values and using 10-fold cross-validation is presented in Table 4. The MEC showed that the model explained between 30 to 59% of clay and OC variation, while the MEC for pH ranged from 50 to 70%, revealing a greater prediction accuracy. The MEC values for clay and pH in the case of Burkina Faso indicated poor predictions (18% and 15%, respectively). When combining the dataset for three countries, the model's performance generally improved, with the highest accuracy observed for the combination of Ethiopia, Nigeria, and Kenya for all three properties. The ME values of all soil properties showed that these were negligibly small compared to the RMSE, indicating that systematic prediction errors were substantially smaller than random prediction errors.

The trained models for each experiment and soil property were used to obtain dissimilarity index maps by AOA. Here, we only present results for OC for two cases: (1) Ethiopia as a donor country; (2) Kenya, Burkina Faso, and Nigeria are the donor while Ethiopia is the recipient country (Fig. 2). Results for other experiments and for clay and pH are provided in the Supplementary Materials. Results indicate that the East-African countries are more comparable to one another because they have lower dissimilarity indices when another East-African country is a donor; also, the West-African countries (Nigeria and Burkina Faso) are in more general agreement based on the dissimilarity index. This behaviour was observed for all soil properties (Section 2 in Supplementary Materials). This was confirmed by studying the spatial average of each DI map (Table 5), which shows that if Kenya is the donor and Ethiopia the recipient, or vice versa, the spatial average DI is relatively small. The same applies to Nigeria and Burkina Faso. For instance, when Ethiopia is the donor, the spatial average DI for OC is 0.38 in Ethiopia and 0.57 in Kenya, while for Nigeria and Burkina Faso the DI averages are 1.17 and 1.33, respectively. In addition, when Burkina Faso is the donor and other countries are the recipients, the DI is large for all properties, most prominently for pH (Figure SM-11 in Supplementary Materials). When Burkina Faso is the donor, the spatial average DI for pH in Burkina Faso is 0.24, whereas this value increases to 4.16 for Nigeria, to 6.72 for Kenya and to 8.77 for Ethiopia (Table 5). This shows that AOA dissimilarity in other countries is large when Burkina Faso is the donor country.

Table 4
Cross-validation results of trained models for the three soil properties and eight experiments. The abbreviations for the mentioned countries are as follows: KE = Kenya, NI = Nigeria, ET = Ethiopia, BF = Burkina Faso.

Experiments	Clay			OC			pH		
	ME	RMSE	MEC	ME	RMSE	MEC	ME	RMSE	MEC
KE	−0.04	16.20	0.31	0.18	15.29	0.37	0.01	0.64	0.72
NI	0.26	12.22	0.57	0.06	6.00	0.43	−0.00	0.55	0.53
ET	0.10	14.30	0.32	0.17	16.26	0.53	0.01	0.64	0.67
BF	0.40	13.44	0.18	0.16	4.80	0.44	0.00	0.58	0.15
KE + ET + BF	0.09	14.37	0.39	0.21	14.52	0.53	0.01	0.63	0.67
KE + ET + NI	0.20	13.70	0.53	0.18	13.11	0.53	0.01	0.60	0.71
KE + BF + NI	0.29	13.26	0.52	0.02	9.23	0.42	−0.00	0.57	0.66
BF + ET + NI	0.22	13.32	0.50	0.27	11.56	0.59	0.01	0.59	0.67

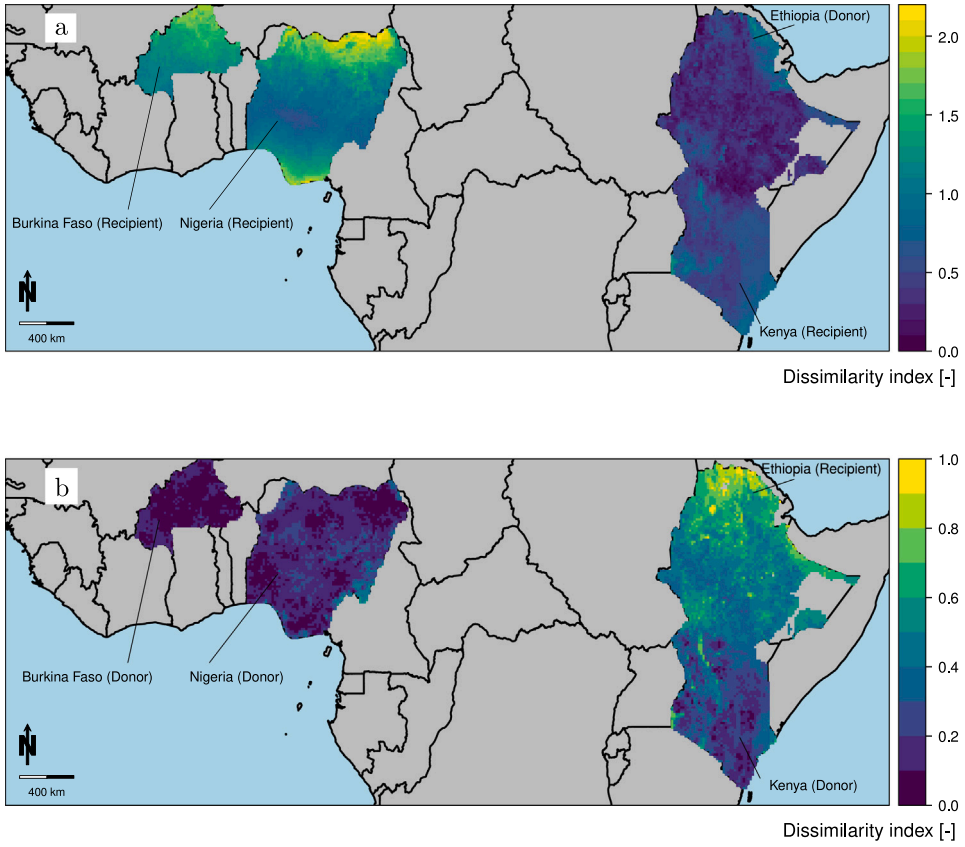


Fig. 2. Dissimilarity index maps of OC; (a) Ethiopia is the donor country and the other countries are recipients, (b) Kenya, Nigeria and Burkina Faso are the donor and Ethiopia is the recipient.

According to the DI maps (Section 2 in Supplementary Materials) and the DI spatial averages (Table 5), combining three countries as donors results in a decline in the mean and range of the DI in all experiments. Fig. 2.b shows a case where Burkina Faso, Nigeria, and Kenya are the donor countries and Ethiopia is the recipient country. The DI map of Ethiopia shows considerable spatial variation and, in general, a high dissimilarity, especially in the northern part of the country.

Fig. 3 shows the density distributions of the DI of the donor country/ies versus the DI of the recipient countries. There is some coverage between Ethiopia (donor) and Kenya (recipient), but not much with Nigeria (recipient), and nearly no overlap with Burkina Faso (recipient) (Fig. 3.a). The DI distribution in the case of Burkina Faso as a donor country is narrower compared to others, whereas the DI distribution of the recipient countries are quite flat (e.g. Figure SM-31, page 15 in Supplementary Materials), meaning that the covariates in Burkina Faso are different from the covariates in the other countries.

When the model is trained on data from three countries, the overlapping of DI plots between the donors and recipient country in all experiments expanded and the dissimilarity decreased. This is visible in Table 5 where the spatial average DI remarkably reduced by combining the training datasets of three countries.

3.4. Uncertainty and comparison

Maps of uncertainty estimates were produced by deriving 90% prediction intervals using the QRF approach. Here also, as mentioned in Section 3.3, we only present figures for two experiments in the case of OC (Fig. 4); other maps are provided in the Supplementary Materials. Although there were differences in the level of uncertainty, all experiments generally showed a similar spatial pattern of uncertainty between countries from the same region (Section 4 in Supplementary Materials). Ethiopia's PIW map for OC (Fig. 4.a) showed some small pockets of high uncertainty in the eastern parts of the country, while the whole country had narrow prediction intervals and performed well in recipient countries, except for Burkina Faso. This is confirmed by Table 5, where the mean PIW values for Ethiopia, Kenya, Nigeria, and Burkina Faso are 44, 59, 57, and 84 g kg^{-1} respectively.

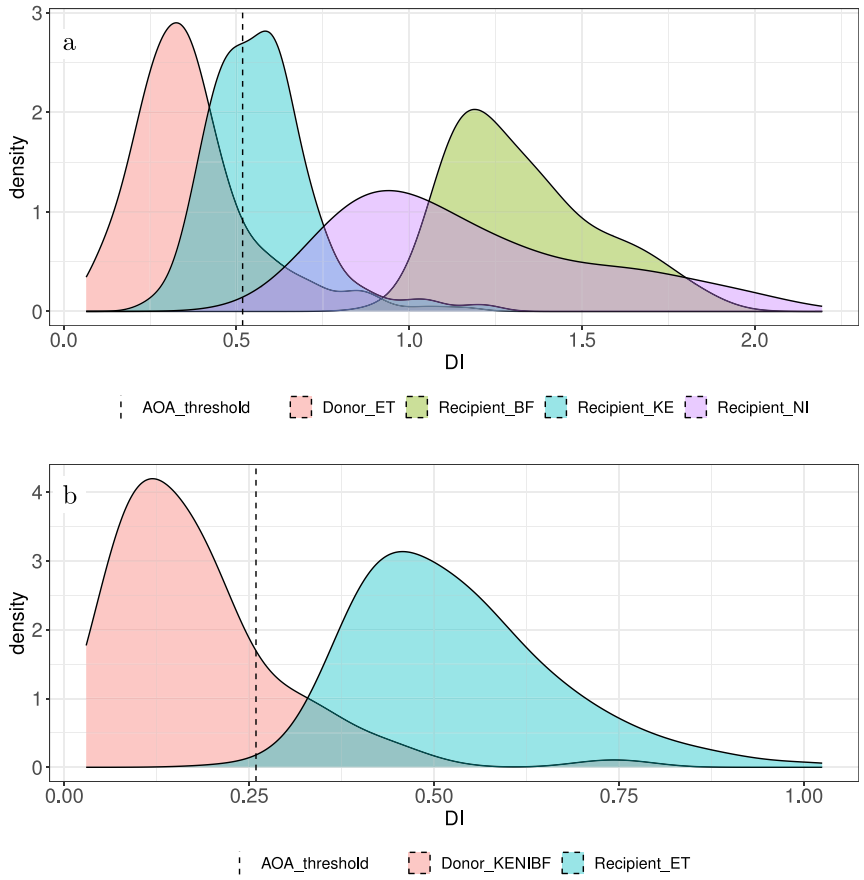


Fig. 3. Distribution of the DI for OC; (a) Ethiopia is the donor and other countries are recipients, (b) Kenya, Nigeria and Burkina Faso are the donor and Ethiopia is the recipient.

Table 5
Spatial average of dissimilarity index and 90% Prediction Interval width.

Donor	Recipient	DI			PIW		
		Cl	OC	pH	Cl	OC	pH
KE	NI	2.28	2.57	0.94	57.63	36.57	3.55
ET	NI	1.04	1.17	0.90	54.04	57.03	3.13
BF	NI	3.92	1.90	4.16	49.62	22.19	2.23
NI	NI	0.15	0.21	0.18	36.34	16.83	2.05
KEETBF	NI	0.46	0.49	0.44	54.65	31.25	2.91
KE	KE	0.28	0.30	0.27	50.81	21.80	2.8
ET	KE	0.45	0.57	0.43	51.09	59.90	3.48
BF	KE	4.47	3.12	6.72	51.84	25.21	2.71
NI	KE	0.79	1.12	1.06	61.49	30.00	3.14
ETNIBF	KE	0.34	0.36	0.34	53.36	51.76	3.46
KE	ET	0.64	0.78	0.48	57.51	34.02	3.28
ET	ET	0.29	0.38	0.24	49.57	44.90	2.74
BF	ET	5.73	3.53	8.77	52.12	30.26	2.60
NI	ET	0.88	1.18	1.10	57.76	31.69	3.08
KENIBF	ET	0.46	0.53	0.45	57.78	33.38	3.18
KE	BF	3.42	3.82	1.09	56.16	21.21	3.46
ET	BF	1.12	1.33	0.81	50.74	84.95	3.02
BF	BF	0.26	0.20	0.24	39.95	11.87	1.95
NI	BF	0.47	0.52	0.47	58.97	17.35	2.63
KEETNI	BF	0.30	0.38	0.29	55.21	25.76	3.08
KEETBF	KEETBF	0.23	0.27	0.19	49.70	32.44	2.76
ETNIBF	ETNIBF	0.17	0.17	0.13	44.71	29.09	2.42
KENIBF	KENIBF	0.19	0.18	0.15	41.81	18.86	2.37
KEETNI	KEETNI	0.18	0.24	0.16	46.90	29.47	2.61

In contrast, the PIW map of Nigeria for pH revealed a high prediction performance in the country itself with the exception of some large portions in the north (Figure SM-52 in Supplementary Materials), but the pH model trained on Nigeria data performed extremely poorly in the recipient countries, as shown in Table 5. The difference between the 0.05- and 0.95-quantile Burkina Faso (donor) maps was large, especially for clay (Figure SM-53 in Supplementary Materials), indicating that the prediction uncertainty was large for the country itself and for other countries.

Using datasets of three countries to evaluate the prediction uncertainty in a recipient country revealed that extrapolation was associated with high uncertainty. In some cases, the uncertainty was high in some parts of the donor countries as well. For instance, the PIW map of clay, when Kenya is the recipient and other countries are donors, showed not only wide PIW for Kenya but also high uncertainty for Ethiopia (Figure SM-58).

3.5. Statistical validation of random forest models

Results of the statistical validation of the models that were trained in donor areas and applied to recipient areas are presented in Table 6. Overall, the statistics in this table – when compared to those presented in Table 4 – confirm that extrapolation leads to larger prediction errors. In fact, in most experiments, the MEC values were negative or close to zero, meaning that the RF model performed worse than using the average of all measurements in the recipient country as a prediction. However, it is easier to predict for neighbouring countries, as for example shown for Ethiopia and Kenya where the spatial prediction of pH had a MEC of 23% (Ethiopia as the donor) and 22% (Kenya as the donor). It is also interesting to note that ME values are sometimes quite large compared to RMSE values. This shows that extrapolation can lead to systematic prediction errors of similar magnitude as random errors, which is rarely the case for interpolation (e.g. see Table 4). For example, when Ethiopia serves as a donor and Nigeria is the recipient, the ME and RMSE values for OC are 11.49 and 13.63 g kg^{-1},

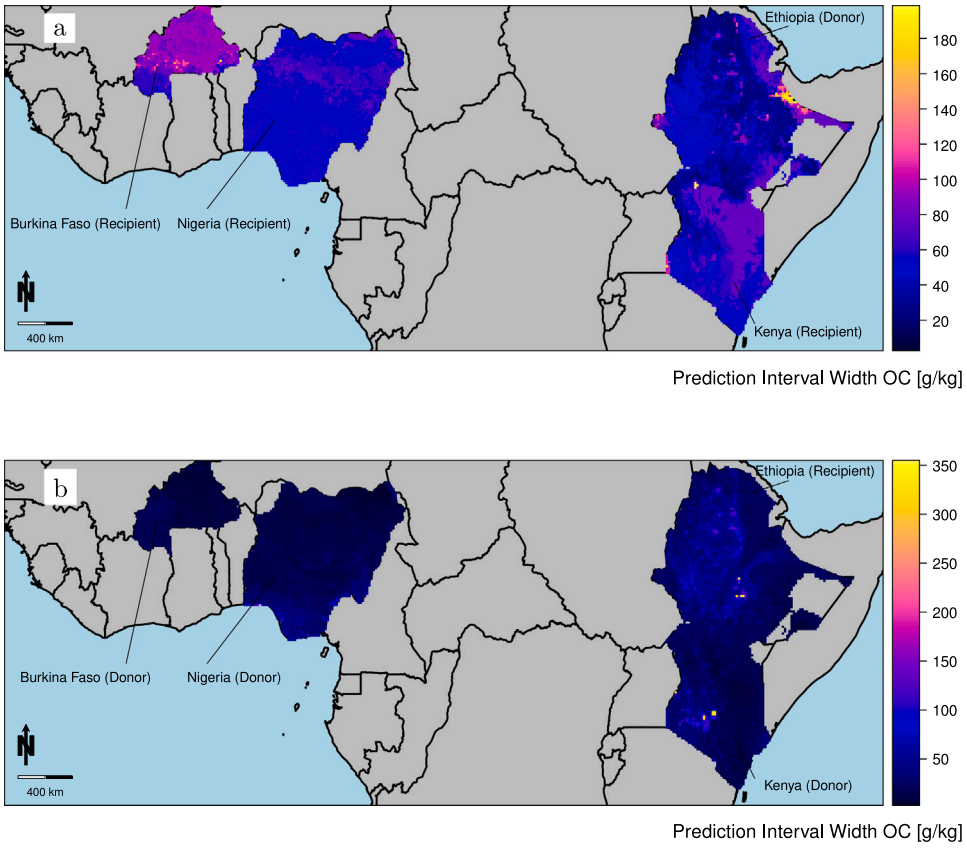


Fig. 4. Prediction interval width of OC; (a) Ethiopia is a donor country and other countries are recipients, (b) Kenya, Nigeria and Burkina Faso are donors, and Ethiopia is the recipient. Note the different scales.

Table 6
Final validation metrics; comparison between all predictions.

Donor	Recipient	Clay			OC			pH		
		ME	RMSE	MEC	ME	RMSE	MEC	ME	RMSE	MEC
KE	NI	8.08	18.51	0.02	1.63	6.96	0.24	0.71	1.01	−0.61
ET	NI	17.87	27.52	−1.16	11.49	13.63	−1.92	0.45	0.89	−0.26
BF	NI	3.00	18.08	0.07	1.61	7.60	0.09	0.42	0.89	−0.24
KEETBF	NI	7.04	19.74	−0.10	2.15	8.16	0.03	0.19	0.70	0.22
ET	KE	−2.20	18.24	0.13	6.77	18.19	0.11	−0.36	1.06	0.23
BF	KE	−15.54	24.82	−0.61	−2.94	18.23	0.10	−0.61	1.33	−0.22
NI	KE	−13.92	23.39	−0.43	−3.66	18.66	0.06	−1.08	1.50	−0.54
ETNIBF	KE	−5.93	18.57	0.10	8.52	18.75	0.05	−0.41	1.02	0.28
KE	ET	2.09	16.66	0.07	−7.55	25.18	−0.14	−0.38	0.98	0.22
BF	ET	−12.37	21.43	−0.54	−11.33	26.31	−0.24	−0.71	1.31	−0.40
NI	ET	−9.17	19.56	−0.28	−14.01	27.44	−0.35	−1.11	1.51	−0.86
KENIBF	ET	0.83	17.05	0.03	−7.51	25.08	−0.13	−0.58	1.08	0.05
KE	BF	4.55	15.40	−0.08	−1.52	6.07	0.11	1.01	1.23	−2.79
ET	BF	12.34	19.50	−0.74	5.95	8.34	−0.69	0.95	1.16	−2.39
NI	BF	1.34	14.25	0.07	−1.68	6.23	0.06	0.05	0.63	0.00
KEETNI	BF	3.56	15.00	−0.03	−0.59	6.08	0.10	0.37	0.74	−0.37

respectively, indicating that in this case systematic error is dominant over random error. The results also showed that training the model for three countries to predict in the fourth is not performing much better than using data from only one country. The only clear exception is predicting pH in Nigeria, which had a substantially larger MEC when the RF model was trained on data from the three other countries.

The plots of Fig. 5 show the relationship between the different measures of extrapolation and the results of the statistical validation metrics for OC. Similar scatter plots for clay and pH are provided in the Supplementary Materials, Figures SM-67 and SM-68. In these plots we used the similarity values of the soil type and homosoil approaches (Tables 2 and 3), while the DI and 90% PIW dissimilarities (Table 5) were multiplied by −1 to achieve that bigger values mean higher similarity in all four cases. Next all metrics were separately linearly re-scaled between 0 and 100, so that the lowest and highest value for each approach were 0 and 100, respectively. We expected a positive correlation of these metrics with MEC and a negative correlation with RMSE (Table 7) but the results did not confirm this. None of the measures of extrapolation had a strong correlation with the validation metrics. From the four measures of extrapolation, only soil type and homosoil approaches exhibited some correlation with the validation metrics, with a slightly stronger correlation observed for soil type. The MEC — soil type correlations were consistently above 0.33 across all three properties, while homosoil yielded the highest correlation with

Table 7
Pearson correlation coefficients for the four investigated methods to estimate the potential for extrapolation (columns) versus the validation metrics (rows). PIW: mean prediction interval width; DI: mean dissimilarity index. Both PIW and DI are multiplied by −1 for consistent correlation interpretations.

Validation metric	variable	-PIW	-DI	Homosoil	Soil type similarity	Soil type similarity with taxonomic distance
MEC	Cl	−0.19	0.25	0.49	0.50	0.25
MEC	OC	0.51	−0.09	0.21	0.36	0.12
MEC	pH	0.15	−0.02	0.43	0.33	0.07
RMSE	Cl	0.18	−0.26	−0.27	−0.29	−0.31
RMSE	OC	−0.03	0.05	0.08	−0.25	0.14
RMSE	pH	−0.14	−0.37	−0.31	−0.42	−0.18

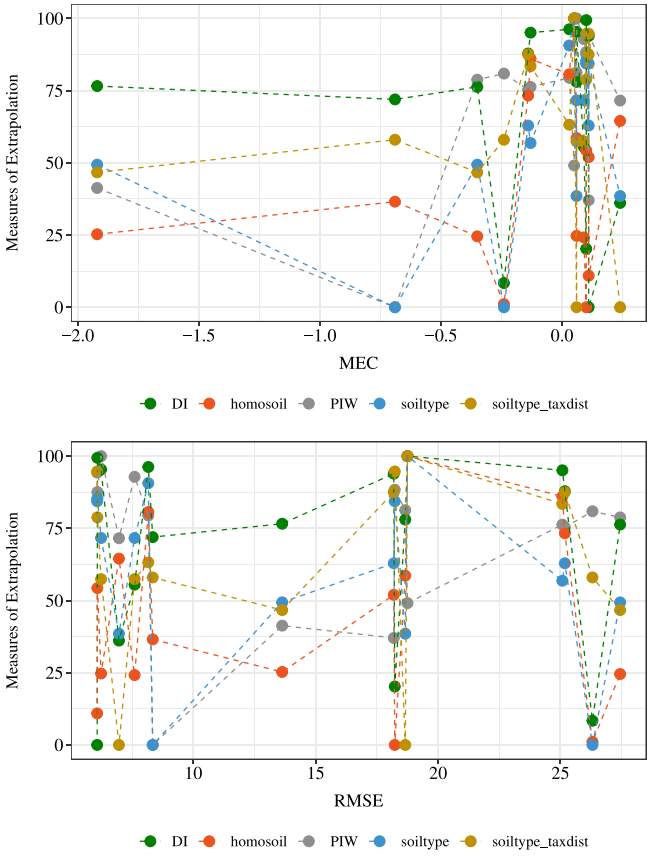


Fig. 5. Relationship between final validation metric and different measures of extrapolation for OC. PIW and DI were multiplied by −1 to ensure that a bigger number means higher similarity. Values on *y*-axis were linearly re-scaled to a minimum of 0 and a maximum of 100 for all five cases.

MEC for clay and pH. Surprisingly, the measures of PIW, DI, and soil type with accounting for taxonomic distances showed low correlations with the validation metrics. In particular, PIW demonstrated a negative correlation (−0.19) with MEC for clay, whereas a relatively high positive correlation (0.51) was observed for OC.

4. Discussion

4.1. Different measures of similarities

The soil characteristics of Kenya, Ethiopia, and those of Nigeria to a lesser degree, were more heterogeneous than those of Burkina Faso. This is due to several factors, such as variations in climate (comprising of different climatic zones as mentioned in 2.1), topography (differences in elevation, Fig. 1), composition and number of soil types (reflecting differences in the number of soil types per countries, SM-1). This finding may explain why these countries exhibit greater similarities with other countries in terms of soil type and the homosoil approach. A larger soil heterogeneity in a donor area can have a significant impact on the prediction capability of a trained model. For example, we found that the trained models for Kenya and Ethiopia were more effective at transferring the model to other areas. This is also supported by the dissimilarity maps and plots generated using the AOA method, which showed smaller DI values when Kenya or Ethiopia were used as a donor country (Section 2 in Supplementary Materials). Furthermore, using three donor countries instead of one improved model performance by increasing the heterogeneity of the data.

We found that similarities in terms of all measures of extrapolation are relatively higher between countries from the same region (e.g. East Africa or West Africa), meaning that they also have more similar soils. Our findings indicate that geographical proximity has a significant impact on the ability of a model to be extrapolated to recipient areas, as confirmed by Nenkam et al. (2022) and Angelini et al. (2020).

The correlations between four extrapolation measures discussed in the study – homosoil, soil type similarity, dissimilarity index by AOA, and QRF prediction interval width – is provided in the Supplementary Materials (Figure SM-69). As expected, a positive correlation was found between homosoil, soil type and DI. However, to our surprise, the correlations between homosoil and DI with PIW were negative. To further illustrate this point, consider the comparison between Figs. 2.a and 4.a. In Fig. 2.a, where Ethiopia is the donor, Kenya exhibits the lowest dissimilarity, while Burkina Faso and Nigeria have the highest dissimilarity with Ethiopia. However, when we compare these results with Fig. 4.a, it becomes apparent that, according to the PIW, Nigeria has the smallest prediction interval width. This suggests that extrapolation based on a model that was trained on Ethiopia might not be so challenging for Nigeria. Kenya, which appeared to be the easiest to extrapolate based on Fig. 2.a, demonstrates more difficulty in extrapolation according to Fig. 4.a. In other words, Nigeria transitions from being ranked third in terms of dissimilarity in Fig. 2.a to being ranked first for easier extrapolation in Fig. 4.a in terms of PIW. These results confirm a negative correlation between DI and PIW. This result contradicts previous findings by Malone et al. (2016), who reported that areas with high dissimilarity typically exhibit greater prediction uncertainty compared to areas with low dissimilarity. It should be noted that the negative correlation between DI and PIW is based on comparing country averages, which is not the same as a comparison on pixel level. According to Simpson's paradox (Norton and Divine, 2015) or the ecological fallacy principle (Freedman, 1999), different results might be obtained depending on the aggregation level. Such analysis was however beyond the scope of this research. The negative correlation between DI and PIW can be further evaluated by considering Table 6. When Ethiopia is a donor and Nigeria a recipient, the RMSE for OC is 13.63 g kg^{-1}, whereas the RMSE increases to 18.18 g kg^{-1} when Ethiopia is a donor and Kenya a recipient.

4.2. Extrapolation results

The cross-validation results for the trained model were deemed acceptable, as seen in Table 4. However, during extrapolation, as indicated in Table 6, the models performed quite poorly. Overall, the RMSE values were high and the MEC values mostly negative, highlighting the potential danger of extrapolation. In Nigeria, the RMSE for training the model for OC is 6 g kg^{-1} (Table 4), but when extrapolating to Burkina Faso, the RMSE increased to 6.23. Furthermore, using the trained model from Nigeria in Kenya and Ethiopia results in a notable decrease in accuracy, with RMSE values of 18.66 and 27.44 g kg^{-1}, respectively. Although the application of soil data using three donor countries decreased dissimilarity and prediction interval width, the validation results indicated that it had only a limited effect on improving extrapolation.

The validation results showed that prediction in a recipient country is more difficult than prediction in the donor country because extrapolation in geographic space often goes together with extrapolation in feature space, and clearly it is more difficult to predict outside the feature space covered by the training data. We found a positive relation between geographic distance and DI; however, there are some notable exceptions. For instance, in Figures SM-6 to SM-8, Burkina Faso and the eastern part of Kenya exhibit a lower DI than the western part of Kenya if Nigeria is the donor country. A similar spatial pattern emerges when Burkina Faso is the donor country (Figures SM-9 to SM-11), although the DI values are scaled differently. In another example, we can see noticeable differences in patterns between the southern and northern parts of Nigeria as recipient in Figures SM-3 and SM-4, despite being at an almost equal distance from Kenya, the donor country.

One possible cause of low performance in extrapolation might arise from the choice of model. RF has been acknowledged as the most proven model in several DSM studies due to its capability of dealing with complex and non-linear relationships between predictors and response variables (Hengl et al., 2015). Furthermore, RF has been demonstrated to be effective in prediction of a soil property of interest when a sufficient amount of training data are available. However, RF, like many other statistical models, faces the challenge of extrapolating in feature space, which limits its application when there are significant areas without observations or when new covariates exhibit distinct characteristics from those learned by the trained model (Meyer and Pebesma, 2021). In other words, RF cannot make predictions beyond the data range, meaning it cannot make predictions that are larger than the maximum or smaller than the minimum observed in the training dataset.

The models generated for soil pH exhibited higher accuracy compared to clay and OC (Table 4) and performed slightly better in terms of extrapolation (Table 6). The reason for the better performance of soil pH may be attributed to its stronger relationship with covariates. In fact, Dharumarajan et al. (2022) and Nenkam et al. (2022) also demonstrated better performance of soil pH in DSM.

Comparing our results with others, Grinand et al. (2008) found that the predictive accuracy was limited when the trained model was extrapolated to another area. Nenkam et al. (2022) also noted that transferring the model between areas which are considered 'homosoil' in relation to each other resulted in weak performance, despite homosoil being a potent tool for transferring soil properties between areas. The study conducted by Malone et al. (2016) revealed that the degree of similarity based on their homosoil approach (slightly different from the approach used in our research) between the regions was approximately 47%, revealing the limited capacity for extrapolation.

4.3. Correlation between extrapolation measures and validation metrics

In addressing the second objective, overall there was a positive correlation between the extrapolation measures and MEC and a negative correlation with RMSE, which is as expected. But the correlations were quite small and not practically significant (as shown in Fig. 5 and Table 7). We were particularly surprised to find weak correlations between validation metrics and both PIW and DI. We had expected them to show stronger correlations than homosoil and soil types due to their reliance on training data, covariates, and calibrated models. Considering that homosoil and soil type similarity exhibit a stronger correlation with validation metrics and are much simpler to compute, one may question the justification for the additional effort to employ PIW and DI in such analyses. Note that homosoil and soil type methods do not require training data and covariates in both donor and recipient areas, nor the fitting of a machine learning model. Our comparison of PIW and DI results,regarding both objectives, revealed that neither method provides a reliable assessment of the quality of maps in extrapolation and it raises questions about the added value of PIW and DI for assessing extrapolation risks. This is a very surprising result that calls for more studies to check if this is a structural phenomenon. If our findings are confirmed by other studies, this is important for researchers who use PIW and DI (as well as AOA) to assess the suitability of models for extrapolation and to delineate areas where predictions are valid and where not.

Our study suggests that blindly using the AOA is not recommended, as we found almost no correlation between AOA (that is, DI) and RMSE and MEC. This contradicts the findings of Meyer and Pebesma (2021), who proposed that information about AOA can be a useful tool to indicate the quality of predictions when applying a model to a new environment. Another study by Ludwig et al. (2023) suggests that creating global maps that are useful for future applications requires restricting predictions to the area of applicability of the model. Although their findings highlight the need for caution when applying machine learning to make predictions beyond the range of covariate values used during model training, our experiments conducted in different countries with diverse environmental conditions showed that the knowledge of DI derived from AOA has little correlation with the final validation metrics. Despite these opposing perspectives, all of these studies contribute to the ongoing development and refinement of spatial extrapolation measures.

4.4. Weaknesses and limitations

This study has identified some limitations that should be considered in future research. One limitation of our study is that it represents only a small set of experiments. We recommend conducting more studies to evaluate whether the poor relation between extrapolation measures and validation metrics is persistent across a wide range of applications.

Another potential limitation is the utilisation of the Jaccard method to compute taxonomic distance, as discussed in Section 2.4.1. It is worth noting that the Jaccard method is just one of several approaches that could have been employed. It should be noted that there is currently no widely accepted method for comparing soil type composition between different countries, which led us to develop our own method. While this approach allowed us to obtain valuable insights, it is important to acknowledge that it involved some subjective decisions. Therefore, further research could aim to establish a more standardised approach for comparing soil type composition between different regions.

Also, another limitation relates to the quality of the training dataset, which may be prone to errors, ranging from field sample collection to laboratory measurements, as the soil surveys were done by different researches in different periods and analysed in different laboratories, often using different methods. This might have contributed to the sometimes large Mean Error statistics that we observed, which are not captured by any of the four extrapolation metrics that we computed.

Finally, the reliability of our validation metrics could be a limitation. However, the experiments that we did had fairly large datasets used for validation based on measurements at locations that were fairly uniformly distributed across the recipient area. Therefore, while this

limitation should be acknowledged, we are not overly concerned about its impact on our conclusions.

5. Conclusion

This study aimed to investigate and compare different measures of extrapolation – including soil type, the homosoil approach, dissimilarity index by AOA and QRF prediction interval width – to determine the potential of extrapolating a machine learning soil prediction models in geographic space. We have reached the following conclusions based on the results and discussion:

- Between different measures of extrapolation, a positive correlation was observed between homosoil, soil type and DI, as expected.
- Contrary to our initial expectations, a negative correlation was observed between homosoil and PIW, as well as between DI and PIW.
- Employing the trained model from a donor country to make predictions in three recipient countries, the extrapolation results demonstrated poor performance. Specifically, the predicted results exhibited an increase in RMSE and ME, while a decrease was observed in MEC.
- Using three donor countries instead of one led to improvements in soil type similarity, homosoil score, accuracy of the trained model, as well as a reduction in dissimilarity and PIW. However, the results indicated that training the model with data from three countries did not yield better predictions in the recipient country compared to using data from only one country.
- None of the four presented extrapolation measures showed a strong correlation with the final validation metrics.
- Soil type and homosoil methods demonstrated a stronger correlation with validation metrics in comparison to DI and PIW, which is quite surprising. In this study soil type and homosoil served as better indicators of extrapolation capability. These methods can be computed before collecting data and training a model. This makes them attractive to explore the extrapolation risk.
- DI and PIW were found to be inadequate measures for assessing the quality of extrapolation in recipient areas in this study. Their inability to provide a reliable indication of extrapolation feasibility raises questions about their continued use for this purpose in general. More research is needed to evaluate if this result is confirmed in other studies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors highly appreciate the support of Andree M. Nenkam and Alexandre M.J.-C. Wadoux for providing the environmental covariates used in this study. We thank Johan Leenaars and Jetse Stoorvogel for expert advice on the physiography and soils of the four African countries.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geoderma.2023.116740.

References

Afshar, F.A., Ayoubi, S., Jafari, A., 2018. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. Geoderma 315, 36–48. http://dx.doi.org/10.1016/j.geoderma.2017.11.030.

Angelini, M.E., Kempen, B., Heuvelink, G., Temme, A.J., Ransom, M.D., 2020. Extrapolation of a structural equation model for digital soil mapping. Geoderma 367, 114226. http://dx.doi.org/10.1016/j.geoderma.2020.114226.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de Forges, A.C., Morgan, C.L., . . ., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Reg. 20, e00255. http://dx.doi.org/10.1016/j.geodrs.2020.e00255.

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A.B., 2014. GlobalSoilMap: Basis of the Global Spatial Soil Information System: Basis of the Global Spatial Soil Information System. CRC Press, http://dx.doi.org/10.1201/b16500.

Asgari, N., Ayoubi, S., Jafari, A., Demattê, J.A., 2020. Incorporating environmental variables, remote and proximal sensing data for digital soil mapping of USDA soil great groups. Int. J. Remote Sens. 41 (19), 7624–7648. http://dx.doi.org/10.1080/01431161.2020.1763506.

Awad, M., Khanna, R., 2015. Efficient Learning Machines : Theories, Concepts, and Applications for Engineers and System Designers. Apress Open, New York, http://dx.doi.org/10.1007/978-1-4302-5990-9.

Batjes, N.H., Ribeiro, E., Van Oostrum, A., 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019)ardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). Earth Syst. Sci. Data 12 (1), 299–320. http://dx.doi.org/10.5194/essd-12-299-2020.

Booth, T.H., Nix, H.A., Hutchinson, M.F., Busby, J.R., 1987. Grid matching: a new method for homoclime analysis. Agricult. Forest Meteorol. 39 (2–3), 241–255. http://dx.doi.org/10.1016/0168-1923(87)90041-4.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. http://dx.doi.org/10.1007/BF00058655.

Brenning, A., Bangs, D., Becker, M., Schratz, P., Polakowski, F., 2018. Package 'RSAGA'. In: The Comprehensive R Archive Network. https://CRAN.R-project.org/package=RSAGA.

Bui, E.N., Moran, J.C., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australiaurray–darling basin of Australia. Geoderma 111 (1–2), 21–44. http://dx.doi.org/10.1016/S0016-7061(02)00238-0.

Dharumarajan, S., Lalitha, M., Niranjana, K., Hegde, R., 2022. Evaluation of digital soil mapping approach for predicting soil fertility parameters—a case study from karnataka plateau, India. Arab. J. Geosci. 15 (5), 386. http://dx.doi.org/10.1007/s12517-022-09629-8.

Du, L., McCarty, G.W., Li, X., Rabenhorst, M.C., Wang, Q., Lee, S., . . ., Zou, Z., 2021. Spatial extrapolation of topographic models for mapping soil organic carbon using local samples. Geoderma 404, 115290. http://dx.doi.org/10.1016/j.geoderma.2021.115290.

Freedman, A.D., 1999. Ecological inference and the ecological fallacy. Int. Encycl. Soc. Behav. Sci. 6 (4027–4030), 1–7, Retrieved from https://web.stanford.edu/class/ed260/freedman549.pdf.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial contextscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma 143 (1–2), 180–190. http://dx.doi.org/10.1016/j.geoderma.2007.11.004.

Hateffard, F., Novák, T.J., 2021. Soil Sampling Design Optimization by using Conditioned Latin Hypercube Samplingatin Hypercube Sampling. Tech. Rep., Copernicus Meetings, http://dx.doi.org/10.5194/ismc2021-35.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., . . ., et al., 2015. Mapping soil properties of africa at 250 m resolution: Random forests significantly improve current predictionsfrica at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10 (6), e0125814. http://dx.doi.org/10.1371/journal.pone.0125814.

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

Jenny, H., 1994. Factors of Soil Formation: A System of Quantitative Pedology. Courier Corporation.

Jones, A., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Dewitte, O., . . ., et al., 2013. Soil Atlas of Africa. European Commission.

Kinoshita, R., Roupsard, O., Chevallier, T., Albrecht, A., Taugourdeau, S., Ahmed, Z., van Es, H.M., 2016. Large topsoil organic carbon variability is controlled by andisol properties and effectively assessed by VNIR spectroscopy in a coffee agroforestry system of costa rica. Geoderma 262, 254–265. http://dx.doi.org/10.1016/j.geoderma.2015.08.026.

Lagacherie, P., McBratney, A., 2006. Chapter 1 spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping, Vol. 31. Elsevier, pp. 3–22. http://dx.doi.org/10.1016/S0166-2481(06)31001-X, Retrieved from https://www.sciencedirect.com/science/article/pii/S016624810631001X.

Leenaars, J., Van Oostrum, A., Gonzalez, M.R., 2013. Africa Soil Profiles Database, Version 1.1. a Compilation of Georeferenced and Standardised Legacy Soil Profile Data for Sub-Saharan Africa (with Dataset). ISRIC Report 3.

Leenaars, J.G., Van Oostrum, A., Ruiperez Gonzalez, M., 2014. Africa Soil Profiles Database Version 1.2. a Compilation of Georeferenced and Standardized Legacy Soil Profile Data for Sub-Saharan Africa (with Dataset). Wageningen: ISRIC Report 2014/01; 2014, ISRIC—World Soil Information: Wageningen, The Netherlands.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and random forest modelsom forest models. Geoderma 170, 70–79. http://dx.doi.org/10.1016/j.geoderma.2011.10.010.

Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., Meyer, H., 2023. Assessing and improving the transferability of current global spatial prediction models. Global Ecol. Biogeogr. http://dx.doi.org/10.1111/geb.13635.

Mallavan, B., Minasny, B., McBratney, A., 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe. In: Digital Soil Mapping. Springer, pp. 137–150. http://dx.doi.org/10.1007/978-90-481-8863-5_12.

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. http://dx.doi.org/10.1016/j.geoderma.2015.08.037.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52. http://dx.doi.org/10.1016/S0016-7061(03)00223-4.

Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (6), 983–999, https://www.scopus.com/record/display.uri?eid=2-s2.0-33745174860&origin=inward.

Meyer, H., Milà, C., Ludwig, M., 2023. CAST: 'caret' applications for spatial-temporal models.: 'caret' applications for spatial–temporal models. Retrieved from https://CRAN.R-project.org/package=CAST. R package version 0.7.1.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633. http://dx.doi.org/10.1111/2041-210X.13650.

Minasny, B., Fiantis, D., Mulyanto, B., Sulaeman, Y., Widyatmanti, W., 2020. Global soil science research collaboration in the 21st century: time to end helicopter research. Geoderma 373, 114299. http://dx.doi.org/10.1016/j.geoderma.2020.114299.

Minasny, B., McBratney, A., 2010. Conditioned latin hypercube sampling for calibrating soil sensor data to soil properties. In: Proximal Soil Sensing. Springer, pp. 111–119. http://dx.doi.org/10.1007/978-90-481-8859-8_9.

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance, and the world reference base. Geoderma 155 (3–4), 132–139. http://dx.doi.org/10.1016/j.geoderma.2009.04.024, (132).

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles — A discussion of principles. J. Hydrol. 10 (3), 282–290. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

Nenkam, A.M., Wadoux, A.M.C., Minasny, B., McBratney, A.B., Traore, P.C., Falconier, G.N., Whitbread, A.M., 2022. Using homosoils for quantitative extrapolation of soil mapping models. Eur. J. Soil Sci. e13285. http://dx.doi.org/10.1111/ejss.13285.

Neyestani, M., Sarmadian, F., Jafari, A., Keshavarzi, A., Sharififar, A., 2021. Digital mapping of soil classes using spatial extrapolation with imbalanced data. Geoderma Reg. 26, e00422. http://dx.doi.org/10.1016/j.geodrs.2021.e00422.

Ng, W., Minasny, B., Malone, B., Filippi, P., 2018. In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. PeerJ 6, e5722. http://dx.doi.org/10.7717/peerj.5722.

Norton, H.J., Divine, G., 2015. Simpson's paradox. . .and how to avoid it. Significance 12 (4), 40–43. http://dx.doi.org/10.1111/j.1740-9713.2015.00844.x.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012. European soil data centre: Response to European policy support and public data requirements. Land Use Policy 29 (2), 329–338. http://dx.doi.org/10.1016/j.landusepol.2011.07.003.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240. http://dx.doi.org/10.5194/soil-7-217-2021.

R Core Team, R., et al., 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria, Retrieved from https://www.R-project.org.

Tan, K.H., 1995. Soil Sampling, Preparation, and Analysis. CRC Press, http://dx.doi.org/10.1201/9781482274769.

Vaysse, K., Lagacherie, P., 2015. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in languedoc-roussillon (France). Geoderma Reg. 4, 20–30. http://dx.doi.org/10.1016/j.geodrs.2014.11.003.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. http://dx.doi.org/10.1016/j.geoderma.2016.12.017.

Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B., 2019. Sampling design optimization for soil mapping with random forestom forest. Geoderma 355, 113913. http://dx.doi.org/10.1016/j.geoderma.2019.113913.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A.X., Hann, S., Burt, J.E., Qi, F., 2011. Updating conventional soil maps through digital soil mapping. Soil Sci. Am. J. 75 (3), 1044–1053. http://dx.doi.org/10.2136/sssaj2010.0002.

Zhang, H., Zhang, R., Qi, F., Liu, X., Niu, Y., Fan, Z., . . ., et al., 2018. The CSLE model based soil erosion prediction: Comparisons of sampling density and extrapolation method at the county level model based soil erosion prediction: Comparisons of sampling density and extrapolation method at the county level. Catena 165, 465–472. http://dx.doi.org/10.1016/j.catena.2018.02.007.

Zhang, H., Zimmerman, J., Nettleton, D., Nordman, D.J., 2019. Random forest prediction intervalsom forest prediction intervals. Amer. Statist. http://dx.doi.org/10.1080/00031305.2019.1585288.